# čeština doma a ve světě

## ROČNÍK IX

ÚSTAV
ČESKÉHO
JAZYKA
A TEORIE
KOMUNIKACE
FILOZOFICKÉ
FAKULTY
UNIVERZITY
KARLOVY

Vydává čtyřikrát ročně
Filosofická fakulta UK v Praze, centrum pro DeskTop publishing
(technické zpracování Drahomíra Marešová)
v nákladu 400 výtisků

Redakční rada:
Jaroslav Bartošek, Naďa Bayerová, František Fröhlich,
Milan Jelínek, Zdena Palková, Vladimír Smetáček, Alexandr Stich,
Dušan Šlosar, Ivan Vyskočil, Jiří Žáček

Redigují členové Ústavu českého jazyka a teorie komunikace FF UK:
ROBERT ADAM (výkonný redaktor), Ivana Bozděchová, Dagmar Čapková,
Jiří Homoláč, Eva Hošnová, Andrea Hudáková, Radomila Kotková,
Ondřej Koupil, Karel Kučera, Alena Macurová, Petr Mareš, Iva Nebeská,
Jiří Rejzek, Lucie Římalová, Karel Šebesta, Jasňa Šlédrová,
OLDŘICH ULIČNÝ (vedoucí redaktor), Irena Vaňková

Toto dvojčíslo vzniklo v Ústavu Českého národního korpusu FF UK.
Redaktor čísla: MICHAL ŠULC

Grafické zpracování Peter Balhar

Adresa redakce:
ÚČJTK FF UK, nám. J. Palacha 2, 116 38 Praha 1
Telefon: (02) 216 19 210

ISSN 1210-9339

Cena výtisku je 60 Kč + 5% DPH

Objednávky vyřizuje
ediční oddělení FF UK, nám. J. Palacha 2, 116 38 Praha 1,
tel. (02) 216 19 298

Podávání novinových zásilek
povolila Česká pošta, s. p., odštěpný závod Praha,
čj. nov. 5423/95 ze dne 26. 9. 1995

# ČEŠTINA DOMA A VE SVĚTĚ

## Téma čísla: Korpusová lingvistika a čeština

Původem latinské slovo *corpus* k nám přišlo z angličtiny, kde znamená jednak sbírku všech psaných textů jednoho druhu nebo jednoho člověka, jednak sbírku informací či materiálů určených ke studiu. Dnes znamená ve světové lingvistice (a v počeštěné podobě také v lingvistice české) soubor počítačově uložených textů, cíleně shromážděný pro vědecké studium jazyka a pro elektronické zpracování jazykových nástrojů a příruček. V tomto čísle ČDS se nám představí stomiliónový reprezentativní korpus psané synchronní češtiny SYN2000, který pro veřejnost zpřístupnil Ústav Českého národního korpusu, a některé další součásti projektu nazvaného možná neskromně, ale snad plným právem Český národní korpus (ČNK). Články se nutně omezují jen na některé vybrané okruhy a problémy, zřetelně však dokumentují význam tohoto nového typu referenčního zdroje. Představují jak šíři jeho záběru (zejména v kapitole *Pohled do korpusu*), tak mimořádnou možnost výzkumů frekvenčních (citovaná čísla se vztahují k frekvenci jevů v korpusu SYN2000, jehož skladba je následující: próza 11,02 %, lit. faktu 2,60 %, poezie 0,81 %, drama 0,21 %, jiné imaginativní texty 0,36 %, životní styl 5,55 %, technika 4,61 %, sociální vědy 3,67 %, vědy o umění 3,48 %, přírodní vědy 3,37 %, ekonomie a řízení 2,27 %, právo a bezpečnost 0,82 %, víra a náboženství 0,74 %, administrativa 0,49 %, publicistika 60 %).

Tím větší je vděk za dar, čím méně se s ním otálelo.
Seneca (O dobrodiních)

# Na úvod poděkování a vzpomínka

Na počátku 90. let, kdy jsme začali uvažovat o tom, že vybudujeme český korpus, a seznamovali jsme se se situací ve světě, začali jsme hledat i domácí spojence. Obraceli jsme se dopisy na různé instituce, v článcích pro různé noviny a časopisy jsme oslovovali potenciální sponzory, snažili jsme se vysvětlit naši situaci a nalézt podporu.

Nalézt ji však nebylo snadné. Zejména na začátku. Každá nová instituce, každý nový počin musí nejdříve proniknout do povědomí lidí, musí se snažit o pochopení a sympatie. Lidské představy o tom, co je a co není důležité, mají velkou tendenci k setrvačnosti a neměnnosti. S tímto faktem musel bojovat a bojuje i náš projekt. Vyvolával a ještě stále vyvolává u některých lidí nedůvěru.

Sponzoři se nakonec našli a my jim děkujeme na jiných místech. Ale náš skutečně nejčestnější sponzor je paní Marie Hlisnikovská, která nám poslala na úplném počátku 300 Kčs. V té době už byla důchodkyně a bylo to pro ni jistě víc než pro všechny instituce, které nám později daly další a větší sponzorské dary.

Kdo to je nebo byl paní Marie Hlisnikovská? V telefonním seznamu je několik Hlisnikovských, dvě ženy toho jména bydlí na Praze 6, stejně jako kdysi paní Marie. Na prvním čísle zvedla telefon paní Jana. „Ne, Marie Hlisnikovská na tomhle čísle nebydlí, ale byla to moje babička, nejlepší babička pod sluncem."

Paní Marie Hlisnikovská byla celý život v domácnosti a bydlela na Praze 6. Uměla pět jazyků. Vydala básničky v lužické srbštině, byla členkou Klubu přátel Lužice. A naše jediná soukromá sponzorka. Zemřela před dvěma roky.

Díky, paní Marie Hlisnikovská, za Vaši pomoc. Už Vám kytici růží nemůžeme donést, alespoň Vám ji dáváme symbolicky.

*Věra Schmiedtová*

# POHLED DO KORPUSU

## Island

V prvním odstavci následujícího textu je číselným indexem naznačena frekvence názvů některých přírodních úkazů. Je patrné, že čím jsou u nás zmíněné přírodní úkazy obvyklejší, tím mají vyšší frekvenci. Výjimkou je pouze slovo *poušť*, které se v češtině velmi často užívá přeneseně. Ve druhém odstavci je naopak označena frekvence některých českých rčení.

Zavítá-li český člověk na Island a poznává-li po nějakou dobu islandskou krajinu, její skály[3970] spojené s úskalími je zdolat, její kopce[3471] pokryté kopami sněhu, její štěrkové pustiny[357] a písečné pouště[2083], její mohutné vodopády[583], z nichž na každého padá posvátná hrůza, její rozlehlá jezera[3463] a prudké říční jezy[454], úbočí zjizvená potoky[2447], její sopky[636], jedny sípající v bahně[1033] a jiné sypající z kráterů[464] lávu[265], její ledovce[967], táhnoucí se bez ladu a skladu, její gejzíry[189], s grácií stoupající k modrému nebi..., jistě si vzpomene na některá česká rčení a úsloví.

Především se mu vybaví dětská hra na hledání předmětů, při níž se hledajícímu pomáhá slovy: „Samá voda[37] — samá voda — přihořívá — hoří!!!" Nebude v tomto případě dalek představy Island'ana, který po dlouhé cestě mezi řekami, jezery a vodopády po výbuchu sopky hledá v horké lávě ztracené předměty. Stejně tak se českému návštěvníkovi náhle do nových souvislostí dostane věta „Ať se propadnu[16], jestli...". Jen na Islandu lze poznat pravý dosah takového sdělení. Stačí se totiž postavit do blízkosti bublající bahenní sopky a propadnutí je téměř zajištěno. Teprve po návštěvě Islandu si také Čech dokáže jasně představit, co to je, když někdo v rozčilení dští síru[9] nebo když se ocitne na horké půdě[62]. Naopak cesta po ledovci může českého cestovatele ovlivnit natolik, že po návratu své problémy bude řešit s ledovým klidem[46] a starosti přímo odloží k ledu[13]. Ale možná také, že z něj budou cestou tryskat celé gejzíry nápadů[9], po návratu spustí vodopády slov[6] a vrhne se do práce. Nechme to na něm. Český člověk se neztratí.

*Renata Blatná*

# Zdrobněliny v korpusu

Zdrobněliny jsou produktivním slovotvorným typem a pro studium zákonitostí jejich tvoření byl využit korpusový materiál.

V Příruční mluvnici češtiny jsou podstatná jména zdrobnělá charakterizována takto: „Deminutiva (slova zdrobnělá) jsou desubstantivní odvozeniny, jejichž formant je nositelem přídatného významového znaku menšího rozměru a/nebo pragmatického rysu hodnocení, častěji pozitivního než negativního." Deminutivní sufixy pak „modifikují významy substantiv všech rodů, a proto mají sufixy pravidelně tři rodové podoby, které respektují gramatický rod základového substantiva."

Deminutiva jsou tvořena jednak primárními deminutivními sufixy (*-ek, -ík, -ka, -ko*), jednak sekundárními deminutivními sufixy (*-ček, -čka, -čko*).

Frekvence výskytů slov tvořených deminutivními sufixy vypovídají o produktivitě těchto sufixů a jejich užívání v současném jazyce. Hledání zdrobnělin tvořených pomocí primárních sufixů je náročné, neboť tyto sufixy jsou homonymní (např. sufix *-ek* slouží rovněž k pojmenování výsledků děje, např. *výrobek*, sufix *-ka* vyjadřuje přechylování, např. *učitelka, řidička*, při hledání přípony *-ík* se vypisují též slova se sufixem *-ník*), vyhledaná data je tedy nutné dále poloautomaticky a poměrně náročným způsobem zpracovávat.

V Tabulce 1 je uveden seznam deminutiv nalezených v SYN2000 a seřazených podle frekvence.

**Tabulka 1: Deminutiva řazená podle frekvence v SYN2000**

| rod ženský | rod mužský | rod střední |
|---|---|---|
| *branka* (21997) | *lístek* (6993) | *městečko* (3347) |
| *tabulka* (13005) | *kousek* (5958) | *kolečko* (1338) |
| *stránka* (11787) | *domek* (5229) | *vajíčko* (1222) |
| *knížka* (10129) | *žebříček* (5119) | *tričko* (1211) |
| *maminka* (9897) | *obrázek* (4717) | *víčko* (1053) |
| *babička* (5803) | *dědeček* (3124) | *sluníčko* (966) |
| *špička* (5174) | *vzorek* (3095) | *slůvko* (773) |
| *novinka* (4607) | *dárek* (2997) | *slovíčko* (573) |
| *chvilka* (3719) | *stánek* (2763) | *klubíčko* (426) |
| *lavička* (2256) | *můstek* (1929) | |

*ulička* (2164)
*holčička* (1974)
*zahrádka* (1472)
*školka* (1471)
*svíčka* (1313)
*sestřička* (1157)
*skleničká* (1063)
*sklenka* (917)

*prášek* (1801)
*vozík* (1718)
*stolek* (1583)
*balíček* (1522)
*kroužek* (1485)
*plátek* (1315)
*koníček* (1293)
*zlomek* (1217)
*ček* (1120)
*sloupek* (1095)
*ptáček* (1069)

Z tohoto přehledu je vidět, že výrazně převažují deminutiva ženského rodu, u těch s největší frekvencí se však jedná o takové zdrobněliny, jejichž význam je odlišný od významu základového slova. Výrazně se zde projevuje ta skutečnost, že korpus obsahuje v současné době 60 % novinových textů, kde je právě hojný výskyt nejvíce frekventovaných deminutiv: *branka* (ve sportovních rubrikách), *tabulka* a *stránka*. Další velmi frekventované deminutivum *knížka* se vyskytuje v korpusu nejvíce ve spojení s adjektivem *kupónová* (1178 z 10129 *knížek* je *kupónových*).

Mezi zdrobnělinami rodu mužského dominuje *lístek* (6993), který je šestým nejfrekventovanějším deminutivem vůbec. Deminutivum *žebříček* (5119) se rovněž hojně vyskytuje ve sportovních rubrikách. Nejméně zastoupena jsou deminutiva rodu středního, z nich pak nejčastější je *městečko* (3347).

Pozoruhodné jsou také zdrobněliny, které se vyskytují v korpusu pouze jednou. Některé z nich jazykový korektor textového editoru Word (dále „korektor") nezná. V následujícím přehledu jsou pro ilustraci uvedena pouze některá ojedinělá deminutiva: *autobusáček* (řidič autobusu), *jézéďáček*, *maloměšťáček*, *hafáček* (ve významu *psík*), *hloupáček* (zde ve významu *rohlík-loupák*), *hnědáček* (alkoholický nápoj *brownie*), *linkováček* (zde ve významu *štětec*), *mindráček* (malý mindrák), *mrkáček*, *myšáček*, *obchoďáček*, *odnímáček*, *ohniváček*, *osikáček*, *ostruháček*, *paneláček*, *pochoďáček*, *podemráček*, *podvraťáček*, *poláček*, *prdeláček*, *prosťáček*, *puberťáček*, *smíchováček*, *spacáček*, *sporťáček*, *veksláček*, *vypínáček*, *Žiďáček*, *estébáček*, *Malostraňáček*, *Menšináček*, *Moraváček*, *bistříčko*, *blahobytíčko*, *bumbáníčko*, *pohlavíčko*, *svítíčko* (zde ve významu *sluníčko*), *vzrušeníčko*.

Slova, která se vyskytují v ČNK pouze jednou, avšak korektor je zná, jsou např.: *loupáček*, *láček*, *košiláček*, *motáček*, *moudivláček*, *máček*, *naháček*,

*siváček, slaďáček, sněhuláček, tralaláček, čtveráček, vrabčáček, vrtáček, ledajáček, zlepšováček.*

Dvakrát se v korpusu vyskytují následující zdrobněliny, většinu z nich korektor zná: *dobráček, dvojháček, dvojáček, školáček, fešáček, květináček, metráček, motoráček, osobáček, plaváček, sosáček, sporáček, všiváček, větráček, Židáček, šediváček, čecháček.* Neznámá jsou slova: *bavoráček, cajdáček, šikuláček, esenbáček.*

Třikrát se v ČNK vyskytují korektoru známá slova *fráček, kolegáček, louskáček, pekáček, pribináček, rypáček, slimáček, tabáček, věšáček, čermáček,* neznámá jsou pro něj zdrobnělá slova *bufáček, chcípáček, kolejáček* a *tranzistoráček.*

Šestkrát najdeme v korpusu zdrobněliny: *hrbáček, odznáček, plecháček, posměváček, poštěváček, slamáček, slintáček.* Ty už korektor zná všechny.

*Jana Klímová*

# Jazyková animace

Jednou z věcí, které si můžeme na korpusu ověřit, je fakt, že některá podstatná jména mají v češtině velmi silně omezenou spojitelnost. Patří k nim mimo jiné i abstrakta jako **pozornost, strach, rozum** atd. Ve spojení se slovesy můžeme u těchto abstrakt často vyjádřit jen začátek děje, jeho trvání a konec děje. Škála sloves pro začátek a konec děje však může být docela pestrá. Tak např. ***pozornost*** můžeme *upoutat, přivábit, vzbudit, přitáhnout, zaměřit, přilákat, zaostřit* apod., potom ji něčemu *věnujeme* nebo k něčemu *upíráme*, nakonec nezbývá než naši pozornost *odvést, odvrátit, odlákat* apod. ***Strach*** nám někdo nebo něco *nahání*, takže nakonec z něho strach *dostaneme*, potom ho *máme* a musíme ho *překonat* nebo nás ho někdo musí *zbavit*. S ***rozumem*** je to podobné. Rozum nejprve buď sami *dostaneme*, nebo nás k tomu rozumu někdo *přivede*, potom rozum *máme*, a nakonec o něj *přijdeme*, popř. nám nad něčím *zůstává stát*.

Ve frazeologii se však podobným způsobem mohou chovat i konkréta. Korpusové doklady nám ukazují, že už například nevystačíme s frazémem *hodit do koše*, chceme-li dát najevo, že se nehodláme nějakou záležitostí zabývat. Tato záležitost totiž nejprve *patří do koše*, potom ji tam *hodíme* nebo *házíme*, dále ta záležitost *do koše putuje*, až *skončí* docela a jistojistě *v koši*. Máme před

sebou jakási pomyslná políčka animovaného filmu, z nichž na každém je napsáno jedno sloveso popisující pohyb. Zkuste si to představit: *záležitost házím do koše – záležitost putuje do koše – záležitost skončí v koši.* Viděli jste ten pohyb? Ten děj? Ten thriller? Neviděli? Tak ještě jednou:

**házím.....putuje..... končí.....**

A tak bychom mohli pokračovat: někdo vás *hází*, *dává*, *strká*, *cpe* s někým ***do jednoho pytle***, až s ním *skončíte*, *ocitáte se*, *jste* nebo *se cítíte* ***v jednom pytli***! Ale s pytlem si užijeme *pytel legrace*, a to nejen s různými slovesy. Počkáme, až někdo zavolá „Obušku, z pytle ven!", a jsme rádi, protože v tom pytli je skutečně *tma jako v pytli*. A tak tu máme nový jazykový thriller:

**hází mě.....skončím.....ven!!!**

**Poznámka:** Vymyslete další thrillery! Vaše korpusové nápady v našem počítačovém koši rozhodně neskončí!

*Renata Blatná*

Hojnost námětů vede k hojnosti slov.
Cicero (O řečníku)

# Obuv (1688), boty (5068), střevíce (410)

**Nejraděj nosím solidní a pevné polobotky (132) s koženou podešví.** Tuhle větu by mohl říct Karel Čapek, ale ne my, protože nic takového už neexistuje. Polobotky ano, ale skoro nikdy s koženou podešví. A už vůbec ne solidní. A my máme na botách podrážku, ne podešev.

Co nosíme my? Módní a drahé, určené pro sport, ale nošené každodenně jsou *adidasky (12)*, už méně prestižní, přesto stejně kvalitní jsou původně české *prestižky (0)* alias *prestiže (0)*. Českým sportovním botám se říká obecně *botasky (20)*. Kde je konec *keckám (53)*, které se nosily v padesátých a šedesátých letech? Vyšší frekvenci mají i díky slovnímu spojení *mít něčeho plný/é kecky* (řekneme o něčem, co přesahuje naše síly). České prostředí vytvořilo i *polokecky (2)*. Do skupiny *sportovní obuvi (50)* patří i *plátěnky (5)*, ale současně patří i mezi *letní obuv (1)*. Jak všechny tyhle sportovní boty shrnout pod jeden

název? Že by *tenisky (238)*? Ale ne, tenisky nejsou nadřazený pojem pro sportovní obuv.

Co víme o *cvičkách (23)*? Kdysi byly černé, s gumou na nártu. Tak asi vypadaly ty, o kterých vzniklo úsloví *Seká to jako Baťa cvičky*. My po válce narození známe *jarmilky (0)*, bílé plátěné cvičky s gumou přes nárt. Byly součástí spartakiádního kroje s výrobním názvem Jarmila — odtud jejich název. Mezi sportovní boty musíme zařadit i *kopačky (419)*. Vysokou frekvenci mají proto, že může někdo někomu *dát kopačky* (opustit ho). V zimě nesmíme zapomenout na *lyžáky (7)* a *přezkáče (1)*.

Co se hodí na denní nošení? *Vycházková obuv (26)*: *mokasíny (7)*, *lodičky (169)*. Ty ale prosím nejvýše s čtyřcentimetrovým podpatkem. Myslete na svou klenbu! Dále *semišky (6)*, *lakovky (11)*, kdysi byly moderní *balerínky (4)*, *maďarky (2)*. Mladí si obují *martensky (0)*, *gladiátorky (2)*, *korkány (0)* nebo *traktory (0)*.

A když začíná přituhovat, co potom? Nazujeme si *zimní obuv (4)*: *kozačky (51)*, *kotníčkové boty (2)* nebo *sněhule (3)*, které obujeme spíše dětem. Maminka se pyšnila v padesátých letech teplými plstěnými *tatrankami (1)*. Dřív jsme si i u nás mohli obout *galoše (55)*, někdy se jim říkalo přezůvky, ty byly i do deště. Dnes si do rozmáčeného terénu můžeme obout jen *gumáky (37)*, někdy se jim říká *holínky (275)*, ale to jsou obecně vysoké boty kryjící holeň. Vystupují též v úsloví *plete si holínky a hodinky*. Do hor na těžký terén si obujeme *pohorky (29)* nebo *pohory (3)*. Kdo byl dříve chudší, chodil do hor v *pionýrkách (2)*, které dělaly pěkné puchejře! Na trhu byly i kotníčkové „pořádné" boty *farmářky (2)*, chodilo se v nich i ve městech.

V létě nosíme *letní obuv (1)*: *sandály (162)*, *dřeváky (97)*, *pantofle (167)*. Dříve jsme nosili *kristusky (15)*, gumovým letním botám uchyceným jen mezi palcem a vedlejším prstem říkáme *vietnamky (4)*, někdo *žabky (0)*.

Jsou tu i boty určené speciálním skupinám obyvatel.

*Důchodky (0)* — černé plstěné boty do půl lýtek, vpředu na zip, které byly často k vidění u starších lidí a na jistou dobu se staly levnou módní zimní obuví mladé generace. *Capáčky (2)* jsou první botičky. *Válenky (41)* jsou věci historickou, zapadly do ruského folklóru. Vojáci nosí *kanady (52)* nebo *půllitry (8)*. Kanady obouvají i skinheadi, jsou součástí jejich image, společně s *bombrem (1)* (speciální kožená bunda). Baletky nazouvají *piškoty (2)*.

*Společenskou obuv* (1) vynecháme, protože se jazykově nijak nevymyká,

pouze se odlišuje tvarově a jemností zpracování.

Už jste z toho obouvání celí uondaní? Kdepak máte *domácí obuv (20)*, abychom se mohli zout a přezout? Máme tu *bačkory (198)*. Můžou sloužit jako synonymum pro *přezůvky (89)*. Modelovou představou bačkor je kotníčková plstěná obuv zapínaná na nártu na speciální kovovou přezku. Slovo bačkory má vysokou spojitelnost s přídavnými jmény *kostkovaný, teplý* a *sešmajdaný*. Že skutečně zapadly do našeho vědomí, dokazuje metaforické užití: *být bačkora (ovladatelný); něco stojí za starou bačkoru; zaklepat bačkorama, natáhnout bačkory (zemřít)*. Doma si vezměte třeba *ťapky (2)*. Sedněte si u nás a můžeme pěstovat *bačkorovou kulturu* (dívat se na televizi). Doma si můžeme obout i *pantofle* a pánové tím pádem mohou *být pod pantoflem*, tedy pod nadvládou své ženy.

Když už boty dosluhují, změní se v *křusky (9), šlupky (1), škrpály (29)*, dřív se měnily v *křápy (3)*. Ale pozor, křáp (jednotné číslo) znamená věc bezcennou. Těžkým botám říkáme *bufy (1)*.

**Pozn.**: Čísla v závorce znamenají frekvenci v SYN2000. Podle výše frekvence můžete usoudit, v kterých místech SYN2000 dostatečně nepokrývá náš úzus, který je jistě současný.

*Věra Schmiedtová*

Zvyk je takřka druhou přirozeností.
Cicero (O nejvyšším dobru a zlu)

# Korpus a pravopis

Na adaptaci cizích slov do češtiny (ale i českých slov odvozených cizí příponou) nepůsobí jen jazykový systém, ale významné jsou i sociolingvistické faktory. Rychleji se adaptuje slovo, které patří do okruhu slov každodenního života, a důležitou roli hraje také skupina obyvatel, která slovo užívá (např. zájmové a profesní skupiny své termíny skoro nikdy nepočešťují).

Podívejme se, jakou naději na plnou adaptaci do češtiny má přípona *-ismus*. Podle posledního vydání Pravidel českého pravopisu jsou varianty přípony *-izmus /-ismus* uvedeny v tomto pořadí jako rovnocenné. Při hledání v SYN2000 jsme dostali tyto výsledky:

Slova psaná se „s" (-ismus) se vyskytují 58 734krát, psaná se „z" (-izmus) se vyskytují 898krát.

Podívejme se nejdříve na počeštěnou podobu. Z výsledku, který jsme dostali ze SYN2000, odečtěme 16 výskytů přípony *-izmus*, které se nacházejí v článcích o nové pravopisné reformě nebo jsou to ohlasy na ni, jedná se tedy o metajazykové užití. Ve sportovních přílohách deníků je speciální výsledková tabulka Motorizmus a těchto výskytů je 17. Skutečně relevantních výskytů (s dostatečným kontextem) s příponou *-izmus* je momentálně v SYN2000 863.

Toto číslo se skládá z jednotlivých frekvencí: 66 slov se vyskytuje v této variantě jen jednou, 27 slov se vyskytuje dvakrát, 11 slov se vyskytuje třikrát, 7 slov se vyskytuje čtyřikrát, po jednom nebo dvou slovech jsou zastoupeny výskyty 5 – 8. *Realizmus* se vyskytuje 9x. Skutečně relevantní výskyty nad deset jsou u slov: *darwinizmus* (12), *komunizmus* (14), *socializmus* (17), *metabolizmus* (32), *organizmus* (226), *mechanizmus* (250). Zajímavý je výskyt slova *neologizmus*. Z 35 výskytů lze 31 najít v jazykovědném časopise Slovo a slovesnost, zbylé čtyři výskyty jsou sice z publicistiky, ale z článků o jazyce. Psali je také lingvisté.

Jak máme zhodnotit varianty slov s příponou *-izmus*?

Uživatelé pociťují tuto původně řecko-latinskou příponu jako knižní, patřící především do sféry odborného jazyka, slova vytvořená pomocí této přípony patří do oblasti termínů. To znamená, že je užívají poučení uživatelé. Přesto je dovoleno psát je počeštěně. Pokud počet výskytů je jeden, dva nebo tři, nemá cenu je zvažovat. Varianty s vyšší a vysokou frekvencí jsou buď determinologizované, potom úzus připouští jejich počeštěné psaní, nebo je uživatelé již vnímají jako samostatné slovo, které s příponou *-ismus* není asociováno.

Srovnáme-li propastný rozdíl mezi výskyty s příponou *-ismus* a *-izmus*, musíme konstatovat, že při zrovnoprávnění obou pravopisných podob byly vzaty v úvahu jiné než materiálové ohledy. Musíme se ptát, zda byl zkoumán skutečný úzus jazyka. Nezdá se, že se v blízké budoucnosti obě přípony stanou z hlediska užívání skutečně „alespoň" rovnocennými.

**Pozn.**: Jistě nás na závěr potěší vyjádření, která využívají pragmatických rysů této přípony při tvorbě slov lezdoprdelizmus[1] (1) a vlezdoprdelismus[2] a jeho derivační řady (7), jejichž výskyt je v SYN2000 doložen. Jak již jsme řekli výše, tato přípona má knižní charakter. Spojí-li se se stylově nesourodou složkou, vzniká její humornost.

*Věra Schmiedtová*

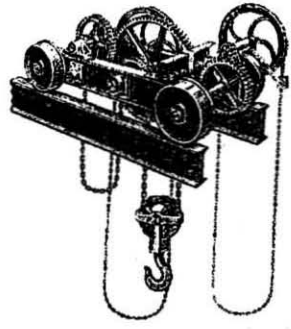
# SLAVNÍ PŘEDCHŮDCI ČESKÉHO NÁRODNÍHO KORPUSU I

Jen málo lidí dnes ví, že o vznik korpusové myšlenky se významně – byť nepřímo – zasloužila nutnost převést soubor rozpadajících se excerpčních lístků Josefa Jungmanna na paměťové médium podstatně trvalejší než papír. Obtížná manipulace s mramorovými excerpčními kvádry, na něž byly po dlouhých sporech během 70. let předminulého století pietně přeneseny výpisky velkého obrozence (podrobněji viz kapitolu *Slepá ulička kvadratické excerpce* v příručce *Manuál lexikografie*), vedla nevyhnutelně k mechanizaci práce s jazykovým materiálem, jejíž výrazný vliv na další vývoj jazykovědy však zůstal bohužel nedoceněn.

Klíčovou úlohu sehrála skutečnost, že konstrukcí mechanismu usnadňujícího práci s mramorovým lexikálním archívem byl pověřen nadšený amatérský filolog V. M. Bíba, jinak strojní inženýr. K jeho zásluhám patří nejen to, že navrhl několik úspěšných zařízení, bez nichž se ani v současné době neobejde žádné lexikografické pracoviště (nejpopulárnější z nich, *Ctirad*, je zachyceno na obr. A), ale zejména to, že si uvědomil nutnost nevytrhávat při slovníkářské práci jednotlivé lexikální jednotky (v Bíbově terminologii „kvádry") z jejich přirozeného okolí, a tedy nutnost pracovat místo izolovaných excerpt s celými texty nebo aspoň s jejich dostatečně velkými úseky. Veden touto myšlenkou sestrojil první v pravém slova smyslu korpusový manažer *Bivoj* (obr. B), který umožňoval prohledávání a třídění celých vět a později (v beta-verzi z roku 1882) i celých odstavců. Cesta korpusové lingvistice se tak otevřela dokořán.

*Karel Kučera*

Obr. A



Obr. B

# Trojúhelníky v korpusu

Není trojúhelník jako trojúhelník. Určitě si nemyslíte, že je to jen obyčejný geometrický rovinný útvar, který má tři strany a tři vrcholy. I takové se samozřejmě liší, neboť můžeme mít trojúhelník rovnoramenný, rovnostranný, pravoúhlý nebo dokonce tupoúhlý. K tvorbě těchto trojúhelníků je třeba znát rozměry úseček, velikosti úhlů, vědět, co je přepona, kde se nachází odvěsna a jak z toho všeho dohromady požadovaný trojúhelník sestavit. Pokud se na to už z geometrie nepamatujete, podívejte se do korpusu – najdete tam návod (bohužel bez obrázku).

Pro svůj tvar a „trojitost" se trojúhelník stal oblíbeným pojmenováním i pro jiné než geometrické skutečnosti. Nacházejí-li se někde tři osoby, místa nebo instituce, pohotově z nich obratný novinář udělá trojúhelník. Ale to mu zpravidla nestačí, protože je třeba ho nějak blíže určit – snad podobně, jako jsme to viděli v geometrii. Tak občas vznikají samostatná pojmenování, která žijí svým vlastním životem. Takovým příkladem je **bermudský trojúhelník**, který známe jako označení oblasti mezi Bermudami, Floridou a Portorikem, kde nevysvětlitelně mizí lodě a letadla. V korpusu na tento jev najdeme celkem 18 dokladů. Další z celkových 44 výskytů se používají přeneseně jako názvy různých „bludných" míst nebo kombinací lidí či institucí, mezi kterými nevysvětlitelně „mizí" dohoda, peníze nebo dokonce rozum a čas, jak čteme v následujícím krásném příkladu z beletrie: *Kráčeli vzhůru k Bermudskému trojúhelníku, jak se v místě říkalo trojcípí hostinců Albína, Záložna, Sport, kde se už po několik generací ztrácel rozum i víra, hodiny i dny...* (Vadný, Josef — Spruzená, Zdenička: Totální brainwash. Pražská imaginace 1990.). Většina zbývajících výskytů se nachází v publicistice a vypadají asi takto: *Česká politika se ztrácí v bermudském trojúhelníku Havel — Zeman — Klaus.* Hodně rozvitým případem pak je *česko-bermudský transformační trojúhelník* (Mladá fronta Dnes 1996). Na závěr této záhady je snad ještě třeba dodat, že texty v korpusu SYN2000 dávají přednost psaní s velkým písmenem (26 výskytů).

Dalším typem trojúhelníků jsou místní označení oblasti, která ho připomíná tvarem (jako vrcholy lze určit tři města) nebo je vklíněna mezi tři státy. Příkladem takového českého trojúhelníku je **lázeňský trojúhelník** (13 výskytů), často s přívlastkem západočeský, jehož vrcholy jsou města Karlovy Vary, Mariánské Lázně a Františkovy Lázně. Dalším místním vymezením je **Zlatý trojúhelník**, což je odlehlé území na hranicích Barmy, Laosu a Thajska, kde se pěstuje opiový mák jako hlavní zdroj příjmů obyvatelstva. Stejně se nazývá (ovšem mnohem

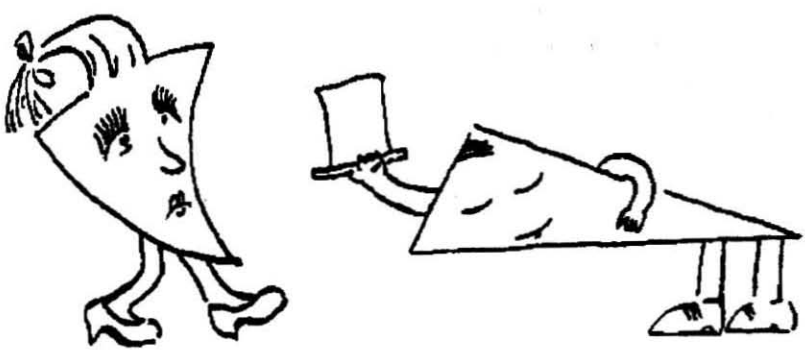
méně často) v roce 1991 speciálně vytvořená zóna volné ekonomiky v Severní Koreji v oblasti Radžin – Sonbong.

Při nahlédnutí do korpusu s hrůzou zjistíme, že se na našem území nachází několik **černých trojúhelníků** (dokonce někdy nazývané i černý trojúhelník Evropy). Jedná se o silně znečištěné oblasti Krušnohoří, Jizersko-Krkonošský trojúhelník (strany trojúhelníků zde tvoří ČR, SRN a Polsko) a dále Třinecký (ČR, Polsko, Slovensko). Z českých trojúhelníků připomeňme ještě areál **Zelený trojúhelník** v Plzni. Jako trojúhelník se také označují seskupení tří států. Zatímco **visegrádský trojúhelník** (tvořený Polskem, Maďarskem a Československem) v roce 1993 rozdělením Československa zanikl (možná se přeměnil na čtverec), v roce 1991 vznikl **výmarský trojúhelník** (Francie, Německo, Polsko). Názvy byly zvoleny podle míst, kde probíhala jednání.

V dopravě se kromě **výstražného trojúhelníku** můžeme setkat s **kolejovým**, **jednostopým** nebo **dvoustopým trojúhelníkem**: *V roce 1967 byla zjednodušena křižovatka Štěpánská na dvoustopý trojúhelník, protože jednostopá trať ve Štěpánské a na Václavském nám. byla zrušena* (Čech, J. – Prošek, F. – Fojtík, P.: Trolejbusy v Praze 1936—1972. Společnost městské hromadné dopravy, Praha 1994.). Z odborné oblasti ještě připomeňme název souhvězdí Trojúhelník.

Výčet „korpusových trojúhelníků" by samozřejmě nebyl úplný, kdybych nezmínila **magický**, **milostný** a **manželský trojúhelník**.

*Marie Kopřivová*



Nemůže dobře mluvit ten, kdo není znalý věci.
Cicero (Brutus)

# Světelný rok

Světelný rok je vzdálenost, kterou urazí světlo ve vzduchoprázdnu za 1 rok. Je to jednotka používaná především, ne-li výhradně, v astronomii a příbuzných vědách. Kdyby se totiž měly vzdálenosti mezi nebeskými tělesy měřit v jednotkách běžných v pozemských měřítkách, muselo by se operovat s obrovskými čísly, což je nepříjemné.

Pro ty, kteří si chtějí udělat představu v běžných jednotkách, se udává přepočet na kilometry. Při rychlosti světla 300 tisíc km/s a délce roku 365 dní je tedy jeden světelný rok 9 460 800 000 000 km.

Nedávno mě děti zaskočily otázkou, zda se náhodou nepočítá přestupný rok. Ne, nepočítá. A i kdyby, rozdíl mezi oběma hodnotami by byl 1/366, tedy asi tři a půl promile. („Přestupný" světelný rok by měřil 9 486 720 000 km.)

Někteří lidé si občas neuvědomují, že jde o jednotku vzdálenosti, nechají se svést slovem „rok" a používají termín světelný rok pro určení časového úseku. Typickým příkladem je jeden verš v písničce s názvem „Zamilovaná", kterou před lety zpívala zpěvačka Lenka Filipová: *ten okamžik trval celý světelný rok*.

Několik pěkných příkladů nepochopení tohoto termínu najdeme i v korpusu SYN2000. Tak například ve Vyškovských novinách mají rubriku Okénko pro děti, kde v článku „Jak čas pořád letí" v roce 1997 vyšlo:*Nic netrvá věčně. Snad jenom čas, který je nekonečný a bdí nad tím, aby na světě mělo všechno svůj začátek i konec. I dny, noci, týdny, měsíce a roky, staletí, tisíciletí a světelné roky mu podléhají. A čas pořád utíká.* Nebo v jednom čísle Reflexu z roku 1994 se můžeme dočíst: *Na rozdíl od vesmíru, který čítá na stamilióny světelných let, má pozemský den jen dvacet čtyři hodiny.*

Ve většině případů se ovšem termínu světelný rok používá správně, tedy pro vyjádření vzdálenosti.

V ČNK se světelný rok vyskytuje celkem 38krát. 15krát je v jednotném čísle, 23x v množném.

V jednotném čísle jde vždy (kromě jediného případu, který je opět příkladem nesmyslného použití: *až světelný rok po pointě*) o vysvětlení pojmu, podobně,

jako jsem uvedla na začátku tohoto příspěvku. Udivující ovšem je, že se zde vyskytuje několik různících se převodních vztahů vůči jednotkám soustavy SI:

jeden světelný rok je 9,28 biliónu kilometrů

jeden světelný rok znamená 9,46 biliónu kilometrů

1 světelný rok = 9,46 pm

Pouze ve druhém příkladu je vztah správně. Zkratka *pm* ve třetím příkladě znamená pikometr, což je $10^{-12}$ m. Autor chtěl zřejmě použít jednotku petametr (Pm), tedy $10^{15}$ m. V tom případě by i třetí vztah byl dobře.

Výrazu *světelné roky* (v množném čísle) 13x předchází číslovka, většinou vyjadřující velké číslo – stovky, tisíce, milióny, dokonce i miliardy. Pokud před ním nestojí číslovka, je termínu použito obrazně. Např. *Lidé na obou stranách mluví stejnou řečí, ale mnohdy by chtěli být navzájem vzdáleni světelné roky.*

Když už jsme u množného čísla, je třeba připomenout, že se tvoří dvěma způsoby, analogicky ke slovu „rok". Kromě světelných roků se tedy používají, dokonce mnohem častěji, světelná léta. V SYN2000 je jich 99, z toho jednou v 1. pádě a jednou v 6. pádě. Ve všech ostatních výskytech (97) jde o 2. pád, kterému v 95 případech předchází číslovka. Většinou tato číslovka vyjadřuje velké číslo, stejně jako u světelných roků.

Zbylé 2 příklady jsou opět jako u tvaru „roků" použity obrazně, tentokrát dokonce v obou případech nesprávně: *není to ale otázka různých generací, ale celých světelných let* (Judith Krautzová: Milenci); *každý člověk, ... , musí na světě něco užitečného vyrobit: pár bot, operaci slepého střeva, ..., tisíce jiných takových věcí, a z hlediska světelných let je věčnost Shakespearova nerozeznatelná od věčnosti páru bot* (Josef Škvorecký: Hovory s Okťabrinou). Můžeme tedy shrnout, že výraz světelný rok je odborný termín, který slouží k vyjádření velkých vzdáleností. Z číslovek, které mu často předcházejí, by se dokonce dalo usuzovat, že i tato velká jednotka je pro vyjádření vesmírných vzdáleností malá. Možná by se mohly zavést analogické termíny světelné století, světelné tisíciletí. Třeba by se dalo vymyslet i něco hezky českého pro milióny a miliardy světelných let.

Astronomické termíny vždy přitahovaly, a dodnes přitahují, básníky a romanopisce. Jen si zkuste vzpomenout, kolikrát jste se v krásné literatuře setkali s Měsícem, nejlépe v úplňku, nebo hvězdami. Tomuto osudu se nevyhnul ani světelný rok, bohužel velmi často doplácí na to, že se skládá ze dvou slov, z nichž rok je také jednotkou. Ovšem pro úplně jinou veličinu.

*Jaroslava Hlaváčová*

# Pokrývky hlavy

Z módního doplňku, kterým *baret (222)* zpočátku byl, se postupně stala záležitost politická. V korpusu nacházíme jeho vysokou spojitelnost se slovy baskický, černý, červený, modrý, zelený. Každá skupina je nejen nositelem jiné barvy baretu, ale i jiného přesvědčení. Módní je *bejzbolová (1)*, *bejsbólová (1)*, *basebalová (10)*, *baseballová (50)* *čepice*, *čapka*, *čepička* nebo *kšiltovka* (64). Vysoký výskyt *beranice (75)* stále není překvapením, ale poměrně slušný výskyt slova *biret (10)* pro mě překvapením je. Nejčastěji má spojení se slovem *kardinálský* a *červený*. *Bouřka (3)* je jen v písničkách Jiřího Suchého.

*Bonbónek (2)*, malý apartní klobouček, je v korpusu jako doklad užití v minulosti. Zato *buřinka (331)* je tam velice často, a to ve spojení s přídavným jménem *černý* a *kouzelný*. Hádejte proč? Odkazuje na Pana Tau, který buřinkou zatočil (také vysoká spojitelnost) a...

Překvapil i *cylindr (163)* s poměrně vysokou frekvencí. Že bude v korpusu ve vysokém výskytu slovo *čepec (194)*, to nijak nepřekvapilo. Ale do výpisu se plete jeden z kravských žaludků. Je doložen i frazém *dostat se pod čepec*. Tyto údaje ale neplatí o *čepici (1582)*, která má časté spojení s přídavnými jmény (*placatý*), s podstatným jménem *kšilt* (*čepice s kšiltem*) a slovesem *smeknout/sundat* (*smekl/sundal si čepici*).

Čepicemi ovšem můžeme někoho i utlouct. Speciální pokrývky hlavy jako *fez (37)*, *florenťák (1)*, *helma (404)*, *čáka (47)* nás unášejí abecedou vpřed. *Jarmulka (15)* se proti očekávání nepojí s přídavným jménem *židovský*. Nenajdeme ovšem synonymum *kipa (0)*.

Dopracovali jsme se k podstatnému jménu *klobouk (3000)*. Nejčastěji je klobouk *černý* nebo *kovbojský*. Ale taky si můžeme něco *strčit za klobouk*. *Kšiltovka (64)* a *kulich (65)* nás provázejí v každodenním životě a tomu odpovídá i frekvence.

Před časem byla módní ženská zimní kukla. Ale dnes se *kukla (299)* pojí především se slovy *černá* a *maskovací*. Tedy je to vybavení policistů a lupičů. *Leninka (0)* doložena v korpusu není, *lodička (3)* a *narážečka (0)* velice málo nebo vůbec ne. Ani *panamák (10)*, *papacha (15)* nejsou nic moc. Skóre se zlepšuje u *placaté čepice (26)*.

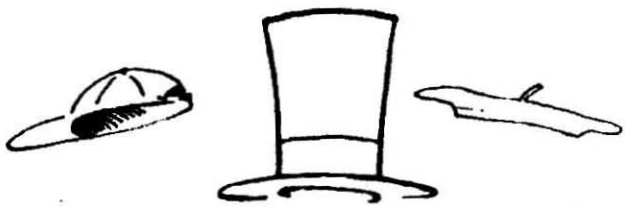
*Přilba (1102)* nabírá na frekvenci. Má spojitelnost hlavně s přídavnými jmény *cyklistická*, *hasičská*, *modrá* (nejvyšší frekvence), *ochranná*. Podobně jako

*helma*. Znamená to, že *helma* a *přilba* jsou synonyma? *Přilbice (18)* nás svou spojitelností s *rytířem* a i jinými výrazy, které se často vyskytují v její blízkosti, odsouvá do historických kontextů. *Rádiovka (59)* a *slamák (61)* vylepšují svou frekvenci, je to důkaz, že se stále nosí. Pokles nastává u *sombréra (19)*. Překvapivě se umístil *stetson (20)*. Na to, že jsem tenhle výraz před svou návštěvou Ameriky vůbec neznala, je pro mě jeho frekvence velmi překvapivá. Zdálo by se, že *sombréro* je mnohem častější. Hle, jak nás naše intuice klame! *Širák (69)* se nám v mnoha variantách spojuje s citátem „svůj širák odhazuji v dál", jak ho známe ze známé hry Cyrano z Bergeraku.

Už finišujeme – *turban (122)* je pokrývka hlavy v některých exotických zemích, ale v češtině si děláme *turban* i v koupelně z ručníku. *Tvrďák (22)* se pojí již jen s divadelním kostýmem nebo s dobou, kdy byl moderní. Skoro stejnou frekvenci má *tvrďák* – muž, který něco vydrží. *Ušanka (34)* se často pojí s přídavným jménem *ruský*, *vydrovka (3)*, *zmijovka (4)* a *žirarďák (0)* nás nesou do cíle abecedy. Jsme v cíli. Nechme už ty *hučky (13)*, *tralaláčky (6)* a *tralaláky (3)* na pokoji! Dejte si na hlavu nějaký *klóbrc (1)*.

Anebo nechte na hlavě.

*Věra Schmiedtová*



## Země slovy korpusu

V korpusových textech se neodráží jen způsob uvažování a vnímání světa, ale také to, jak - coby obyvatelé malé země uprostřed Evropy – vnímáme okolní státy, co nás na nich fascinuje či kam jsme si je „zaškatulkovali". Takovou škatulkou může být pojmenování tvořené spojením **země + genitiv substantiva**. Substantivum v genitivu zde označuje něco pro nás neobvyklého, čím je daná země proslulá, typická nebo zvláštní ve srovnání s naší vlastí. Jsou to například

geografické útvary (jezera, hory, fjordy, gejzíry), stavby (pagody, pyramidy, minarety), výrobky (sýry, špagety), fauna či flóra (sloni, borovice, tulipány), sport (kriket, fotbal, cykloturistika), události z historie země (osvícenství), původní obyvatelé (Inkové) nebo další výrazná specifika.

Mezi země, jejichž metaforické označení najdeme i ve slovníku, patří především **země tulipánů = Nizozemí** (20 výskytů) a **země tisíce (tisíců) jezer = Finsko** (27 výskytů). V korpusu SYN2000 se nově objevuje **země orlů = Albánie** (11 výskytů). Na rozdíl od Nizozemí a Finska, které jsou známé hojností tulipánů a jezer, je země orlů překladem názvu státu z albánštiny. Tyto názvy stojí v kontextu většinou samostatně, např.: *...členství v Evropské unii znamená v tomto ohledu pro zemi tisíců jezer velké pozitivum* (Hospodářské noviny 1995). Všechny další přívlastky jsou méně frekventované a většinou je v kontextu ještě připojeno jméno státu, např.: *V průměru za pětadvacet tisíc lze strávit stejně dlouhou dovolenou na černém kontinentu v zemi faraónů a pyramid, tedy v Egyptě* (Mladá fronta Dnes 1996).

U přívlastků následujících zemí také většinou přesně víme, k jakému státu odkazují, ale v korpusových textech jsou užívány zřídka: **Austrálie** – země protinožců; **Barma** – země (tisíce zlatých) pagod; **Egypt** – země faraónů, pyramid; **Island** – země gejzírů, ohně a ledu, ohně a vody; **Japonsko** – země vycházejícího slunce, země workoholiků; **JAR** – země diamantů; **Kanada** – země dřevorubců, Eskymáků; **Mexiko** – země Indiánů, masek, stínů; **Palestina** – země hojnosti (biblicky), bible; podobně **Izrael** – země otců, předků; **Korea** – země jitřní záře; **Švýcarsko** – země cykloturistiky, čokolády, bank a hodinek, plebiscitů; **Turecko** – země minaretů a bazarů, půlměsíce.

Dále existují přívlastky, pod které se řadí několik zemí: **země sýrů** – Nizozemí, Francie; **země lásky** – Francie, Itálie, Havaj; **země hor** – Anglie, Arménie, Dagestán.

Poslední skupinu tvoří přívlastky, které se vyskytly v jednom díle, pro které je autor vytvořil, aby danou zemi stručně charakterizoval. Někdy je jich více za sebou a nejsou uváděny v ustálené formě. Nejvíce takových přívlastků v korpusu najdeme pro **Ameriku**: země malých měst a předměstských čtvrtí, bezpečnosti (v období druhé světové války), země zaslíbená tenisu, Indiánů, Inků, možností, mrakodrapů i antimrakodrapů, paradoxů, svobody, dolarů, neónů, fotbalu, snů. Zajímavá je i **Francie**: země borovic, bažantů a husích jater, olší a topolů, topolů a platanů, platanů a vinic, osvícenství, lidských práv, azylu. **Dánsko** je pro nás zemí blahobytu, másla a mléka, klidu a pohody, zemí pro spokojené a klidné

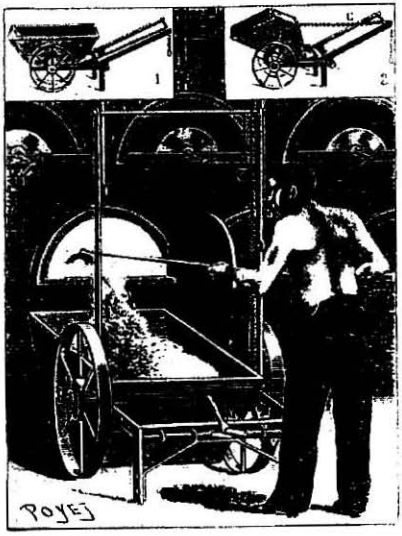
lidi, zemí fjordů (i když toto označení přísluší spíš Norsku). **Itálie** je zemí špaget, **Írán** je země růží a slavíků, **Irsko** země Keltů, **Izrael** země zázraků, rozporů, kontrastů a změn, trpělivosti i násilí, klidu i napětí, země sněhu i vedra, mrtvých skal a písku i bohaté úrody, **Jamajka** země kriketu, **Japonsko** země symbolů, divů a draků, **Keňa** je země běžců, **Německo** země Bacha, Beethovena, Brahmse, **Polsko** země vystěhovalců, ne přistěhovalců, **Rakousko** země dovolené a sousedů, **Rumunsko** je zaslíbená země dobrodruhů, **Španělsko** země toreadorů, **Thajsko** země slonů, nádherných staveb, zlata i chudoby, **Tibet** země sněhu.

Na závěr ještě několik přívlastků pro naši vlast: země husitů a Morgensternů, země hudebníků a textařů, země kritiků (Diviš), lidí bez hrdosti, obrázků Josefa Lady, spálená země Miloše Zemana, země Bojů, plzeňského, extrémů.

*Marie Kopřivová*

# SLAVNÍ PŘEDCHŮDCI ČESKÉHO NÁRODNÍHO KORPUSU II

Prvním korpusovým projektem, v němž bylo plnou měrou využito moderní techniky, se stal moravský nářeční korpus Brněnské borec, přezdívaný též Brněnské drak (příčinu ke vzniku této populární přezdívky zavdala pravděpodobně skutečnost, že korpus byl budován a roku 1891 i zpřístupněn na parním serveru Maryša, který bylo nutno několikrát hasit). Celodenní funkčnost brněnského serveru (zdůrazněme, že šlo o vůbec první parní server na světě) zajišťovala špičková výtopna vybavená unikátním systémem moderních koksových kotlů, navrženým a zkonstruovaným z popudu – a rovněž z prostředků – známého mecenáše korpusové lingvistiky J. Grmely, zámožného opavského pekaře. V horních rozích



dobového vyobrazení korpusové výtopny je dobře vidět rozfázovanou činnost jedné z periférií Maryši, polyfunkčního zařízení umožňujícího operativní přenos jazykových dat na střední vzdálenosti.

*Karel Kučera*

# Kolokací, co hrdlo ráčí

Jak nejlépe v korpusu hledat ustálená spojení, tzv. idiomy či frazémy? Pro výběr těchto spojení ze záplavy kolokací, které nám nabídne korpusový manažer, slouží statistické ukazatele *mi-score* a *t-score*. Ukazatel mi-score (vzájemná informace) umožňuje lexikografům nalézt v korpusu slova, která se spolu vyskytují častěji, než jak ukazuje jejich náhodná distribuce, a zvýrazňuje relativně vzácná idiomatická spojení, která by lexikograf sám obtížně hledal. Ukazatel t-score na druhé straně vyzdvihuje běžná funkční slova a je k prvně jmenovanému komplementární v tom, že zvýrazňuje kontrasty kolokačních kombinací vybraných dvojic slov. Ilustrujme proces zjišťování ustálených spojení na slovese *ráčit*.

V tabulkách 1 a 2 jsou vedle daných slov a statistických ukazatelů uvedeny jednak hodnoty absolutní frekvence, vyjadřující, kolikrát se dané slovo vyskytuje v těsném spojení (zprava nebo zleva) se slovesem *ráčit*, jednak relativní frekvenci, která určuje procento výskytů daného slova těsně u slovesa *ráčit* z celkového počtu výskytů daného slova v korpusu.

**Tabulka 1: Kolokace slovesa ráčit seřazené podle *mi-score***

| slovo | mi-score | t-score | rel. f. | abs. f. |
|---|---|---|---|---|
| 1. milostivě | 13.01 | 2.449 | 5.217 | 6 |
| 2. hrdlo | 12.16 | 7.21 | 2.908 | 52 |
| 3. laskavě | 11.09 | 3.161 | 1.381 | 10 |
| 4. odpustit | 9.755 | 3.737 | 0.5473 | 14 |
| 5. prominout | 9.221 | 2.641 | 0.3782 | 7 |
| 6. tudy | 8.967 | 1.729 | 0.3171 | 3 |
| 7. půvab | 8.796 | 1.728 | 0.2817 | 3 |
| 8. dovolit | 8.624 | 5.086 | 0.2499 | 26 |
| 9. urazit | 8.619 | 1.995 | 0.2491 | 4 |

| | | | | |
|---|---|---|---|---|
| 10. zajisté | 8.432 | 1.727 | 0.2188 | 3 |
| 11. povolat | 8.252 | 1.726 | 0.1932 | 3 |
| 12. dovnitř | 7.544 | 1.989 | 0.1182 | 4 |
| 13. přát | 7.25 | 3.441 | 0.09645 | 12 |
| 14. rozumět | 7.161 | 2.979 | 0.09066 | 9 |
| 15. vstoupit | 6.688 | 2.801 | 0.06533 | 8 |
| 16. zachovat | 6.628 | 2.213 | 0.06265 | 5 |
| 17. dál | 6.557 | 4.079 | 0.05964 | 17 |
| 18. uznat | 6.554 | 1.714 | 0.05951 | 3 |
| 19. prosit | 6.49 | 2.211 | 0.05695 | 5 |
| 20. dotknout | 6.46 | 1.712 | 0.05577 | 3 |

**Tabulka 2: Kolokace slovesa ráčit seřazené podle *t-score***

| slovo | mi-score | t-score | rel. f. | abs. f. |
|---|---|---|---|---|
| 1. se | 3.312 | 10.83 | 0.006293 | 145 |
| 2. " | 3.556 | 9.769 | 0.00745 | 114 |
| 3. , | 1.732 | 8.87 | 0.002105 | 161 |
| 4. hrdlo | 12.16 | 7.21 | 2.908 | 52 |
| 5. . | 1.334 | 6.469 | 0.001597 | 115 |
| 6. já | 3.686 | 5.61 | 0.008155 | 37 |
| 7. dovolit | 8.624 | 5.086 | 0.2499 | 26 |
| 8. být | 1.47 | 4.95 | 0.001755 | 60 |
| 9. ! | 4.692 | 4.509 | 0.01638 | 22 |
| 10. vědět | 5.128 | 4.451 | 0.02215 | 21 |
| 11. dál | 6.557 | 4.079 | 0.05964 | 17 |
| 12. jak | 3.883 | 4.063 | 0.009345 | 19 |
| 13. odpustit | 9.755 | 3.737 | 0.5473 | 14 |
| 14. ty | 4.348 | 3.683 | 0.01291 | 15 |
| 15. ? | 3.182 | 3.446 | 0.005749 | 15 |
| 16. přát si | 7.25 | 3.441 | 0.09645 | 12 |
| 17. vzít | 6.058 | 3.412 | 0.0422 | 12 |
| 18. laskavě | 11.09 | 3.161 | 1.381 | 10 |
| 19. : | 1.834 | 3.136 | 0.002259 | 19 |
| 20. jen | 3.356 | 3.126 | 0.006486 | 12 |

Podíváme-li se v korpusu na kolokace tohoto slovesa seřazené jednak podle vzájemné informace (viz Tabulka 1), jednak podle *t-score* (viz Tabulka 2),

a sečteme-li pořadí jednotlivých kolokátů v obou tabulkách, získáme nejvýznamnější kolokáty slovesa *ráčit* (viz Tabulka 3). Tato slova můžeme na základě korpusových dat považovat za kandidáty na frazeologická spojení slovesa *ráčit* (podrobněji viz Výpisy ze SYN2000).

**Tabulka 3: Pořadí kolokátů slovesa ráčit podle *mi-score* a *t-score***

| slovo | pořadí dle mi-score | pořadí dle t-score | součet pořadí | počet dokumentů | počet autorů |
|---|---|---|---|---|---|
| 1. hrdlo | 2 | 4 | 6 | 41 | 41 |
| 2. dovolit | 7 | 8 | 15 | 10 | 6 |
| 3. odpustit | 4 | 13 | 17 | 9 | 7 |
| 4. laskavě | 3 | 18 | 21 | 7 | 7 |
| 5. dál | 17 | 11 | 28 | 10 | 10 |
| 6. milostivě | 1 | 28 | 29 | 2 | 2 |
| 7. přát si | 13 | 16 | 29 | 10 | 9 |
| 8. prominout | 5 | 26 | 31 | 6 | 6 |
| 9. rozumět | 14 | 22 | 36 | 2 | 2 |

V následujících výpisech z korpusu SYN2000 jsou uvedena některá slova z tabulky 3 a jejich kolokace se slovesem *ráčit*. Důležitým faktorem pro potvrzení hypotézy, že příslušnou kolokaci můžeme považovat za frazém, je vedle statistických ukazatelů rovněž počet jednotlivých dokumentů, ve kterých se dané spojení vyskytuje, resp. počet různých autorů používajících toto spojení.

Kolokaci *co hrdlo ráčí* (viz Výpis 1) můžeme jednoznačně považovat za frazém, neboť se v korpusu vyskytuje 48krát ve více než čtyřech desítkách dokumentů, oba statistické ukazatele mají vysokou hodnotu. Toto spojení použilo 41 různých autorů ve stejném počtu dokumentů.

Spojení *račte dovolit* (viz Výpis 2) a *račte odpustit* je možno na základě dat z korpusu považovat za dosti ustálená spojení, nikoli však za frazémy, statistické ukazatele těchto kolokací mají průměrnou hodnotu a jsou použity dostatečným (nikoli však vysokým) počtem různých autorů. Podobně můžeme posuzovat spojení *račte laskavě, račte dál, račte si přát* a *račte prominout*. Tyto kolokace lze na základě hodnot statistických ukazatelů a počtu autorů používajících toto spojení považovat za poměrně ustálené.

Avšak kolokace *račte milostivě* a *račte rozumět* i přes to, že mají vysoké hodnoty vzájemné informace, za pevná spojení pokládat nemůžeme, neboť byla

použita pouze dvěma různými autory, jejich běžné používání není na základě korpusu prokázáno.

Na závěr tedy konstatujeme, že jediným na základě korpusu doloženým a v současnosti dosud hojně užívaným frazémem obsahujícím sloveso *ráčit* je spojení *co hrdlo ráčí*. Zajímavé je sledovat, co vše může hrdlo ráčit. V korpusu (viz konkordance ve Výpisu 1) nacházíme, že tento frazém se pojí nejčastěji se slovesy *jíst*, *najíst se*, *sníst*, ale také *troubit*, *užívat si*, *klít*, *pít*, *dát si*, *dopřát si*, *konzumovat*, *ochutnat*, *vyhrát*, dále stojí vedle substantiv *víno*, *jídlo*, *pití*, *chleba*, *čaj*, *pivo*, *rumík lékařský*, *zelenina*, *hovězí*, ale můžeme *mít*, *co hrdlo ráčí*, i *kalhot*, *definic*, *svobody*, *zpráv inspekčních* nebo *popela*. Pozoruhodná jsou i posunutá použití tohoto frazému, např. *hrdlo ráčí třeba šest chodů* nebo *co vaše hrdlo ráčí* či *takže jestli hrdlo ráčí, udělej si kafe*.

**Výpis 1: Konkordance spojení *co hrdlo ráčí***

(vzájemná informace je 12.16 a t-score je 7.21, 41 různých dokumentů, 41 autorů)

Je to pravda," smála se Philippa . „Můžu pít, co <hrdlo> ráčí. No, tedy s rozumem." „To je úžasná zpráva!" (Sheperdová, Stella: Vražedný lék. Svoboda, Praha 1995.)

ložnice, toalety, sprchy („horké vody, co <hrdlo> ráčí, uvidíte"). (Crichton, Michael: Koule. Svoboda, Praha 1993.)

Tady se mohli napást co <hrdlo> ráčí. (Ajtmatov, Čingiz: Popraviště. Lidové nakladatelství, 1988.)

V kufru bych vez chlastu, co <hrdlo> ráčí, samý cizí značky. (Ajtmatov, Čingiz: Popraviště. Lidové nakladatelství, 1988.)

kastaněty rozpalovaly krev, hudba i zpěv plné smutku a kvílející vášně, tanec hrdý a divoký, vína co <hrdlo> ráčí. (Shaw, Irwin: Tomu málo, tomu nic. NLN, Praha 1995.)

**Výpis 2: Výběr konkordancí spojení *račte dovolit***

(vzájemná informace je 8.62, t-score je 5.01, 6 různých autorů, 10 dokumentů)

„A <račte dovolit>, s kým mám tu čest?" obrátil se najednou uhrovitý pán na mladíka s uzlem. (Dostojevskij, Fjodor Michajlovič: Idiot. Ikar, 1996.)

„Totiž, <račte dovolit>, jsem pán domu, i když vám neodmítám patřičnou úctu. (Dostojevskij, Fjodor Michajlovič: Idiot. Ikar, 1996.)

„Pane doktore, kdybyste <ráčil dovolit>, dostal jsem nápad. (Zábrana, Jan: Vražda pro štěstí. MFD, 1989.)

„Jestli <račte dovolit>," ozval se pan Poláček, „mně chtěli vyhodit ve fabrice kotelnu. (Zábrana, Jan: Vražda se zárukou. Československý spisovatel, 1989.)

<Račte dovolit>, to je také náhoda, že paní Katuše nám naprosto mizí hned po smrti svého manžela? (Čapek, Karel: Povídky z jedné kapsy. Československý spisovatel, 1993.)

*Jana Klímová*

## Členové rodiny v korpusu

Četný výskyt deminutiva *maminka* v materiálu korpusu byl podnětem pro zjišťování výskytů pojmenování dalších členů rodiny. Vedle deminutivních tvarů uvádím vždy i základní tvar pojmenování rodinného příslušníka.

Bohatost užití zdrobnělin v češtině ukazují frekvence výskytu deminutivních variant označení *matka* (je nejfrekventovanější: 15835 výskytů), resp. slova *máma* (3235 výskytů): 9897krát *maminka*, 609krát *mamka*, 50krát *mamina*, 42krát *mamička*, 9krát *maminečka*.

Hlava rodiny má následující frekvenci: 4549krát *táta*, 4790krát *tatínek*, 149krát *tatíček*, 109krát *tatík*; označení *otec* se používá nejčastěji (19118krát).

U starší generace je nejčastější označení *babička* (5803krát), dále je v korpusu 767krát *babka*, 432krát *bába*, 235krát *prababička*, 224krát *babi*, 35krát *prabába*, 16krát *bábinka* a 11krát *bábuška*. Označení *dědeček* (3124krát) je méně časté než *babička*. *Děda* má frekvenci 921, *dědek* 841, *děd* 805, *pradědeček* 211, *praděd* 145, *dědoušek* 117 a *dědula* 9.

Co se týká ostatních členů rodiny, nalezneme v korpusu 71829krát *dítě* a 468krát *děťátko*, *dcera* má 10553 výskytů, *dcerka* 828 a *dceruška* 253. *Syn* má frekvenci 15936, *synek* 737, *synáček* 295 a *syneček* jen 13.

*Teta* má v korpusu 2282 výskytů, její zdrobněliny *tetka* 73 a *tetička* 346 výskytů. *Strýc* byl použit 1869krát, *strýček* 1034krát a *strejda* 368krát.

Označení *bratr* se v korpusu nachází 11336krát, *bratříček* 272krát a *bratřík* 51krát, *sestra* 9110krát a *sestřička* 1062krát, *sestřenice* 310krát a *bratranec* 861krát.

*Jana Klímová*

## Konference
## Staré známé, nebo neznámé slovo?

Otázka z názvu signalizuje, že se toto slovo změnilo. Ale nakolik a jak? Otázky se dají pochopitelně rozmnožovat dál. Měli bychom se ještě ptát, kolik má toto slovo vlastně významů a jak je chápat, resp. rozlišovat (to bychom se už ptali na nesnadnou otázku polysémie) a nakolik nám stávající slovníky vlastně podávají spolehlivou informaci. Kritizovat minulost je ovšem často snadné a laciné, zvláště když tyto slovníky odrážet současnost nemůžou. Informace z *Českého národního korpusu* (ČNK) nám ukazuje, že tu proti pěkně a jednoduše vykolíkovanému terénu, jak ho vidí slovníky, nic tak jednoduchého není. Slovníky nám říkají, že *konference* je (podle SSJČ):

1-porada (politiků, odborníků ap. k prodiskutování a řešení určitých otázek),
2-porada učitelského sboru,

zatímco (podle novějšího jednosvazkového SSČ) už může mít významů poněkud více:

1-porada politiků, odborníků ap.,
2-pravidelně zasedající orgán: helsinská, biskupská k.,
3-porada učitelského sboru.

Z téměř 4000 kvantifikovaných výskytů slova *konference* (z téměř stomiliónového korpusu) se však rýsuje po velmi hrubé analýze poněkud odlišný obrázek:

1-nespecifický význam porady (ve smyslu 1 výše, 99,363 % všech výskytů),
2-význam sportovní události (vrcholové kolo soutěže, význam nezaznamenaný ani mezi neologismy, 0,485 %),
3-význam porady školní (0,125 %), nebo
4-význam internetové „porady" (otevřená diskusní platforma, dosud nikde nezaznamenaný význam, 0,053 %).

Stojíme tu evidentně před otázkou polysémie téměř *in statu nascendi*, kterou hokejoví či fotbaloví fanoušci či internetoví uživatelé budou ovšem ostře

a právem popírat, protože jim jejich úzus jistě připadá už starý a běžný. Okrajovou důležitost, zatím nerozhodnutelnou, tu má pořadí seřazených významů; význam „školní" se však očividně zatím posunul ke konci řady. Problém rozhodování bychom však museli řešit pouze u významů 2–4, které mají v ČNK zastoupení zcela okrajové. Naproti tomu více než 99 % úzu připadajícího na význam první (kde už nemusí jít jen o poradu odborníků) může vybízet k otázkám zcela jiným, především po příčině tak ohromující převahy a povaze rozšíření tohoto významu.

Klasickým přístupem ke studiu významu bylo v lexikografii už i před Wittgensteinem studium kontextů, a tedy kolokability dané formy, bohužel obvykle bez velkého dalšího rozlišování vnitřních typů; takto se často pro substantivum jednostranně ve slovnících preferuje kolokabilita (spojitelnost) s adjektivem. Podívejme se na kolokabilitu *konference* z hlediska jejích typických a distinktivních kolokací (spojení) včetně slovesných; napřed pro významy řídké nebo nové. Porovnejme napřed typické kontextové příklady delší (KON), popř. větné, a pak distinktivní dvoučlenné kolokace (KOL).

Pro náš druhý význam (výše):
KON...*pořadí východní k. Atlantické divize, prohrál ve finále k.*;
KOL: *zahájit, založit konferenci.*
Pro třetí význam:
KON...*přinést návrh známek na k.*;
KOL: *k. bude, rozhodne.*
Pro čtvrtý význam:
KON...*od svého založení fungují konference*;
KOL: *vytvořit, vyrobit, přihlásit se do konference* aj.

Jakkoliv tu jde o substantivum a jeho vystižení skrze jeho kontexty, vypovídací síla adjektiv je tu pro jeho charakterizaci problematická a velmi malá, např. ve významu 4 a prakticky i 3 se adjektiva téměř neužívají. Obecně to ukazuje na nutnost rozlišovat i typy kolokability a šíře kontextů; pro analýzu úzu a významu je nutné využít, co se nabízí.

Podíváme-li se konečně na nejčastější první význam, vidíme, že tu jsou naopak typické kolokace s verbem:

KOL: *účastnit se, uspořádat, konat k.*, popř. *konference začala, rozhodla, přinesla...*; k substantivním kombinacím lze počítat kolokace jako *delegáti k.*, *zahájení k.*, *konání k.*, *svolání k.*, *k. homosexuálů* apod., přičemž nápadně častým slovem vyskytujícím se ve volném, ne však bezprostředním okolí slova

*konference* je slovo *organizace* naznačující, že konference nebývají spontánní.

Překvapení a výzvu tu však představují adjektivní kolokace. Na jedné straně tu výrazně dominují adjektiva udávající časovou délku nebo druh setkání, srov. např. *dvoudenní* (23x), *celostátní* (39x), *generální* (259x). Na druhé straně však stojí nepřekvapivě *tisková konference*, jejíchž 1194 výskytů představuje absolutně nejčastější kolokaci slova *konference* a jejíž procentuální podíl se blíží jedné třetině výskytů. Tato nejnápadnější skutečnost, dosavadními slovníky nereflektovaná, musí vzbuzovat při tradičním lexikografickém přístupu rozpaky a nelze ji obejít jen pouhým uvedením v exemplifikaci či ocejchováním kolokace jako víceslovného termínu, kterým nesporně je také. Škatulkování typu lemma, příklad, popř. víceslovný termín zde tudíž nestačí. Z jediné kombinace (s adjektivem *tisková*) však nelze na druhé straně ani vyvozovat, že jde o specifický další význam slova *konference*.

Kolokační spektrum tohoto slova v jeho nejčastějším a doloženě i hlavním významu je rozložené do slovnědruhových kombinací (tedy se všemi relevantními autosémantiky):

(1) adjektivních (A-S),
(2) substantivně-verbálních (V-S), řídce pak ještě
(3) substantivně-substantivních (S-S) a
(4) substantivně-verbálních (S-V).

Podíváme-li se konečně na sémantiku *konference* v tomto významu a funkci blíže a pokusíme-li se je vysoudit ze zde nejpříznačnějších adjektivních kolokací, které jsou zde vysoce relevantní, můžeme se i pokusit zpřesnit tradiční popis jejího významu. SSJČ tento její význam vystihuje explicitním poukazem na tyto rysy:

(a) některé obory, (b) problém a (c) diskuze.

Podívejme se napřed na vyskytující se hlavní kolokace v ČNK: *biskupská, celostátní, celosvětová, dvoudenní, generální, helsinská, křemenologická, listopadová, madridská, muzeologická, mimořádná, mírová, obilní, odzbrojovací, postupimská, tisková, třídenní, ustavující, vědecká, vrcholná k.*; ojedinělé další kolokace musely být z důvodů omezení rozsahu ponechány stranou. Odhlédneme-li od zvláštního a terminologického statutu tiskové konference, pak je z předběžné analýzy zřejmé, že běžně a často explicitně určované dimenze, resp. aspekty jsou

(a) profese, obor nebo ad hoc účel (řešený problém),
(b) místo,

(c) čas,
(d) šíře záběru a platnost,

které však z definice SSJČ zřejmé vždy nejsou; aspekty (b) a (c) implicitně přítomné jsou, (c) však často splývá s (a). V dokladech z ČNK je jednak naznačena skutečná šíře zastoupení výše zmíněných „odborníků", jsou tu však ještě další, v definici nezmíněné aspekty. S výjimkami, jako je *biskupská konference*, kde neomezená časová délka signalizuje výjimku, další význam nebo zároveň termín, nám ČNK mj. potvrzuje i pragmatickou zkušenost, že konferenci nelze obvykle pořádat o příliš dlouhém trvání, typicky se ve světě podle ČNK konferuje pouze 2 až 3 dny. Naproti tomu není analogicky běžné v daném sémantickém spektru konferovat kratší dobu, např. jen půl hodiny.

V této souvislosti lze znovu připomenout i věčný teoreticko-praktický problém, kolik vlastně takových dimenzí, aspektů je třeba ke konstituování nového významu; lze přitom pochybovat o tom, zda pro takové určování stačí aspekt jediný, jako např. u některých „významů" výše ap.

*František Čermák*

# Jazyková poradna, korpus a internet (1): O Internetu a internetu

Koncem roku 1999 vyšlo v nakladatelství Fortuna nové, rozšířené školní vydání Pravidel českého pravopisu. Aktualizovány jsou příklady vlastních názvů; do pravopisného slovníku byl doplněn např. *billboard*, *hospic* (-u i -e), *kotace*, *ombudsman*, *scrabble* (tedy relativně ustálené a rozšířené neologismy a slova dříve úzce odborná; celkem přes 100 výrazů). Heslo INTERNET zní:

**internet**, -u m. *(propojené počítačové sítě)*; **Internet** vl. jm. *(Celosvětová informační a komunikační síť)*

Představme si školní dítko, které má napsat, že ve volném čase rádo brouzdá po... Pomůže mu litera Pravidel k volbě malého nebo velkého písmena? A jak dalece pomáhá dospělému čtenáři? Myslím, že by neškodilo pár slov na vysvětlenou.

V jistých čím dál omezenějších případech ještě můžeme chápat *Internet*

jako jedinečný softwarový produkt, psát ho s velkým písmenem a žádat připojení *k Internetu / na Internet*. Velkým písmenem ho přiřadíme k *Passnetu*, *Eunetu*, *Cessnetu* a početným dalším *-netům*. Původně takové názvy označovaly dílčí sítě, a nejen počítačové (*KabelNet*: kabelová televize). Naproti tomu existuje terminologická řada označující počítačové nebo spíš síťové technologie: *internet* (psaný, bez ohledu na jednotné psaní této triády, také Internet) – *intranet* – *extranet*. Dnes se objevuje složka *-net* v názvech firem, které umožňují nákup po síti: říční lodě nabízí *Barges-net*, stavební materiály i v maloobchodním měřítku nabízí *TECHNIservNET*, s. r. o., firma *V-NET – Radek Všetečka* nabízí renovaci tonerů laserových tiskáren a kopírek... (Způsob psaní, zejména kombinování malých písmen s většími a ještě většími, je mnohdy bizarní: i to patří k dnešku.) Dále nalézáme *-net* v adresách internetových kaváren. Kdo se chce seznámit s příslušným výrazivem, může si prohlédnout např. stránku www.adpnet.cz, která mimo jiné obsahuje „slovníček pojmů ze světa internetu, www, sítí". Jazyková úroveň jeho zpracování je nízká: *nástroj pro vzhledávání souborů; doména nejvyšší úrovně užívaná školami; množina příbuzných programů, užívaných na gateway, které ochrňují privátní sítě před externími uživateli; je jednou z technologií pro generování dynamických HTML stránek na straně klienta* a další mikro- i makrolapsy, např. že o Y2K se dovíte, co každý věděl dávno předtím, než tento barnumsky vyhlašovaný problém nenastal, ale nikoli původ tohoto označení, totiž Year 2 Kilo(s), samorostlý výraz pro *rok dva tisíce*. Nicméně orientační představu tam snad získat lze. Ale vraťme se k *i/Internetu*.

Ani počítačoví odborníci se neshodují v tom, zda lze internet chápat jako něčí „produkt", ač právě to je pro velké písmeno zásadní věc. Ale nevadí. Firma Microsoft se postarala o to, že „její" *Internet* nemá konkurenci a stal se celosvětovým informačním **médiem**. Poprávu ho proto píšeme stejně jako *televizi, tisk* aj., tedy *internet*: ostatně firmy nabízející *připojení k internetu*, takzvaní provideři [*provajdři*], už na adrese www.seznam.cz charakterizují své služby právě takto, s malým písmenem. Informační dálnice směřuje tam, kam se dávno před ní odebrala *karma, lux, jekor* a další výrobky: směrem k deproprializaci a malému počátečnímu písmenu. V knihkupecké výloze dnes najdeme vedle sebe příručky *Nebojte se Internetu* a *@beceda internetu*. Ten druhý způsob psaní má nadějnější budoucnost.

Zvykejme si také používat internet pro získávání informací. I o jazykových prostředcích. Na rozdíl od dřívějších generací lexikologů, kteří byli bezmocní bez pracně shromažďovaných excerpčních lístků, máme jako jazykový materiál k dispozici milióny stránek v počítačové síti (nikdo by se ovšem nezlobil, kdyby textů bylo méně a byly kvalitnější) a máme také účinné vyhledávací nástroje

(nikdo by se ovšem nezlobil, kdyby pracovaly jemněji a nezahlcovaly nás balastem).

Jazyková poradna Ústavu pro jazyk český nemůže spoléhat jen na tištěné zdroje. Nemůže ani čekat, až nějaké slovo bude v tisku natolik běžné, že si ho některý lingvista všimne a zpracuje o něm drobnost nebo slovníkové heslo – a do té doby krčit rameny a neradit. Telefonuje nám třeba soudní písařka (představí se: *já píšu vraždy*) a potřebuje vědět, jak napsat název předmětu, jímž byl usmrcen člověk: *tomfa* nebo *tonfa*? Dřevěný předmět leží před písařkou, je to předmět doličný, ale nikterak nedoličuje ani nenapovídá, jak se má psát jeho vlastní označení. Právě díky internetu jsme zjistili, že tonfa, původně zemědělský nástroj, se stal zbraní v asijských bojových technikách (husité s cepy nejsou tedy ve světě sami). Zkrátka měli bychom umět zjistit, ve které oblasti se takové zatím nenormované, nepopsané, nezachycené slovo objevuje, co znamená, jak se psává, a na základě toho je pak posoudit. V dalších částech tohoto miniseriálu si to ukážeme.

*Josef Šimandl*

# Kočkopes

je jedno z módních slov užívaných v současné publicistice. Slovník „Nová slova v češtině. Slovník neologismů" (1998) ho definuje jako expresivní výraz pro něco, „co nemůže tvořit fungující celek, něco nepovedeného". Za definicí je uveden následující příklad: *Ekonomika zřejmě zůstane v podobě* ***kočkopsa*** *na poloviční cestě mezi trhem a centrálním plánováním.* V poznámce je uvedeno, že je to příležitostný publicistický výraz. Prozkoumejme tedy jeho doklady v korpusu SYN2000. V tomto korpusu je zachyceno 46x, z toho pouze 2x v memoárové literatuře a 1x v populárně-naučném textu, zbývající doklady jsou z publicistiky. Převažují doklady z posledních dvou let. Korpus SYN2000 zachycuje toto slovo ve dvou významech. První význam se týká zvířete, které má rysy kočky i psa, druhý význam je přenesený na předměty a jevy, které v sobě zahrnují různorodé prvky.

První význam zachycují např. dva příklady z časopisu Story z roku 1997. První zní takto: *Původně chtěla kočku, ale známí jí to rozmluvili. Tak zvolila*

*kompromis a koupila si boloňského psíka, protože je to takový* ***kočkopes***. Druhý příklad je podobný: *Členem rodiny je i pejsek Japan Chin. Váží dvě a půl kila a jmenuje se Bischop. „Říkám mu kočička milovaná, protože je to spíš ‚****kočkopes****‘.“* V obou případech jde o psa menších rozměrů, který se podobá vzhledem a částečně chováním kočce. Složeninu *kočkopes* bychom proto mohli interpretovat jako ‚kočkovitý (popř. kočkovitě vypadající) pes‘. Bizarnost takového živočicha komentují Lidové noviny z roku 1999: ***Kočkopes*** *by se mohl vyskytnout leda ve Světě zvířat v době, kdy tam byl redaktorem Jaroslav Hašek.* Zcela specifickou představu o zvířeti s tímto názvem měli tvůrci jednoho kresleného seriálu (s případným názvem „Kočkopes“), který vysílala na přelomu tisíciletí televize Nova. Je to jakési siamské dvojče, které má na jedné straně těla hlavu psí a na druhé hlavu kočičí.

Pokusme se nyní prostřednictvím korpusu zjistit, co mají v jazykovém užití kočka a pes společného a čím se liší. Pokud se budeme zabývat nikoli slovy, ale skutečnými zvířaty, učiníme jistě jednoduchý závěr: pes štěká, kočka mňouká, pes žere kosti a hrabe v zemi, kočka dává přednost mléčným výrobkům a šplhá po různých předmětech, tj. v přírodě po stromech, v bytě samozřejmě po skříních. Jejich soužití v jedné domácnosti vypadá asi takto: pes na kočku zavrčí, zaštěká, kočka dotčeně mňoukne, vyleze na knihovnu a pes pobíhá dole a doráží. Tito dva tvorové mají k sobě odjakživa averzi. Co je tedy spojuje? Především to, že jsou to dnes dvě nejčastější domácí zvířata, tedy domácí mazlíčci, které však někteří majitelé vyhodí nebo vyženou a oni se potom toulají, až skončí v útulku – samozřejmě v tom lepším případě. Tento fakt velmi dobře reflektují i jazyková spojení z korpusu SYN2000. Spojení *kočka a pes* (v různých pádech včetně množného čísla) je tu zastoupeno 47x, opačné spojení *pes a kočka* 156x. Psů je v korpusu zhruba 4x víc než koček (12970:3365). Není to proto, že bychom psy měli o tolik radši, ale spíš proto, že je můžeme vycvičit k různým účelům. Tak máme např. *psy bojové, hlídací, horské, honicí, lovecké* (popř. *lovecky upotřebitelné*), *ovčácké, pastevecké, policejní, řeznické, sanitní, slepecké, služební, stavěcí* atd. atd. Psi mohou být *inteligentní, poslušní, mírumilovní, věrní*, nebo naopak *agresivní, nebezpeční, vzteklí* a *zlí*. Pes *štěká, vyje, čuchá, vrtí ocasem, střeží, táhne*, my ho *hladíme*, ale také ho musíme *venčit*, zpravidla je nutno ho *přivázat*, ale navíc ho můžeme na někoho *poštvat*, a když ho máme dost, tak ho *vyženeme*. Naproti tomu *kočka* je převážně *černá, mourovatá, zrzavá, šedobílá, bílá*, dále *vycpaná, nalepená, budoárová, disneyovská*, popř. *bezocasá*, podle rasy *angorská, perská, siamská* atd., někdy dokonce *fantastická, vymyšlená* nebo *mluvící*. Kočka *mňouká, přede, prská, lehce našlapuje, svítí jí oči, mrouská*

*se*, ale pokud ji nechceme, zpravidla ji jdeme *utopit*. Jak psi, tak kočky mohou být *bezprizorní, chcíplí, čistokrevní, divocí, domácí, hladoví, toulaví*, my je *chováme, krmíme, hladíme*, když utečou, *chytáme* je.

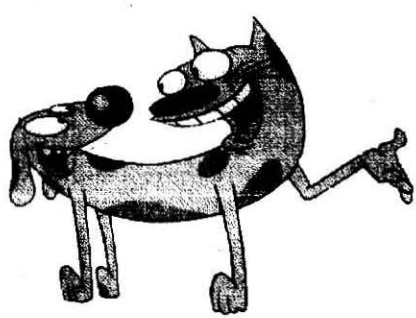
Nyní se stručně vraťme k počátkům užívání slova *kočkopes* v oblasti politické. Dva korpusové doklady vypovídají o jeho původu i možném původci. První z nich je z Reflexu 1994 a zní: *Rozpory v OF, které Macek označil názvem „**kočkopes**", a snahy o diferenciaci politické scény jej nejprve přivádějí do Meziparlamentního klubu demokratické pravice a později mezi zakladatele ODS.* Z toho vyplývá, že slovo bylo v přeneseném významu užito ještě v době existence Občanského fóra, tj. počátkem roku 1992. Druhý doklad, který nepřímo ukazuje na původce tohoto výrazu, je z časopisu Rock&pop z roku 1998: *... posluchač pak nemá dojem, že mu je předkládáno cosi, pro co guru českých konzervativců* (= patrně Miroslav Macek – pozn. aurorky) *vymyslel termín „**kočkopes**"*. V roce 1997 se do souvislosti toto slovo – tentokrát v rámci humorného kontextu – dostává do kontaktu opět se jménem Miroslav Macek: *Rovněž se svěřil* (J. Zieleniec), *že v něm tandem Macek – Švorcová vyvolává dojem **kočkopsa***.

Od roku 1994 se tento výraz aplikuje na různé situace, při nichž dochází k propojení zdánlivě neslučitelných prvků, např. *Neměl by to být nějaký **kočkopes** – mezi socialismem a kapitalismem, ale využití zkušeností obou systémů* (Nedělní příloha Lidových novin, únor 1994). Také nadpis *Zplodili jsme **kočkopsa*** chápe pod tímto zvířetem „lenivý a neschopný státní aparát a zvlčilý, bezohledný soukromý sektor" (Hospodářské noviny 1995). Ještě markantnější jsou doklady z Lidových novin 1997: *Jan Ruml o projevu Václava Havla: „Je to pokus o kvadraturu kruhu, o jakéhosi **kočkopsa**, který v naší zemi existovat nemůže"*. Podobně reaguje na Havlův projev Václav Klaus v časopise Týden: *Václav Klaus dal obratem najevo svou skepsi: „Byl by to opět jakýsi **kočkopes**, který vím, že v naší zemi existovat nemůže."* Seriál uzavírá opět doklad z časopisu Týden: *Shodli jsme se, že bude velmi užitečné, aby do Vánoc jmenování nové vlády bylo skutečností, prohlásil po schůzce Lux, zatímco Klaus zopakoval svůj výrok o „**kočkopsu**"*. Slovo začíná být spojováno s ODS a objevuje se v Mladé frontě Dnes z téhož roku: *Současné tři opoziční strany* (= ČSSD, KSČM a SPR-RSČ, tj. Sládkovi republikáni, pozn. autorky) *jsou totiž velmi heterogenní trojkou a kombinace ČSSD s lidovci je také větší **kočkopes** než stávající trojkoalice*. V magazínu Lidových novin z téhož roku se užití tohoto slova rozšiřuje na představitele Komunistické strany Číny a jejich snahu přiblížit tuto obrovskou zemi k principům tržního hospodaření: *Pohledy na fotu sice*

*směřují vpravo, ale našinec si o* ***kočkopsech*** *myslí svoje...* V roce 1998 proběhly volby a výraz se objevuje u nově vládnoucí strany – ČSSD: *Povolební matematika i názory představitelů některých politických stran v Olomouci naznačují, že by ve městě mohlo skončit čtyřleté vládnutí široké pravolevé koalice, kterou svého času označil někdejší olomoucký zastupitel a nynější ministr kultury Pavel Dostál (ČSSD) za* ***kočkopsa****.* V následujícím roce s odkazem na pokusy o sestavení koaliční vlády, tzv. duhové koalice ČSSD, ODS, KDU-ČSL a US, se tento výraz vrací jako bumerang: *Duhová koalice by byla nefunkčním* ***kočkopsem****, jakýmsi vzkříšeným OF, tedy něčím, proti čemu Václav Klaus zcela správně dosud vedl nesmiřitelný boj. Kočkopes* v politickém smyslu je tedy nefunkční spojení pravicových a levicových proudů v rámci jednoho celku. Nutné je zdůraznit přívlastek nefunkční, neboť s ním souvisejí i další slova a slovní spojení s podobným významem, např. *paskvil, chytrá horákyně, kulatý čtverec, kvadratura kruhu* apod. Výraz *kočkopes* však nenacházíme jen v politické oblasti. Svědčí o tom doklad týkající se zvláštního typu tramvaje: *S finanční podporou spolkového ministerstva pro výzkum a technologie, DB a železničního průmyslu bylo vyvinuto tramvajové dvousystémové vozidlo, v Čechách nazývané „****kočkopes****“.*

Složeninu *kočkopes* dnes zdaleka nelze považovat za ustálenou. Vypovídá o tom odstup autorů, kteří je píšou v uvozovkách (celkem 5 případů), doplňují je zájmeny *jakýsi* a *nějaký* nebo jedno- a víceslovnými přísudky *nazývat, označit názvem* nebo *vymyslet termín*. Další vývoj tohoto kompozita závisí nepochybně na tom, zda se v naší každodenní realitě budou vyskytovat i nadále jevy a situace, které jím bude možné označit.

*Renata Blatná*



Nic není tak nesnadné, aby se to pátráním nedalo zjistit.
Terentius (Sebetrapič)

# Jazyková poradna, korpus a internet (2): Rekupace

Matky lvice jsou v jazykové poradně poměrně častými klientkami, a pokud jde o jejich tvrzení, je namístě obezřetnost: v jejich ústech i fakta jsou především zbraně a podle toho s nimi zacházejí. Předkládám tedy okolnosti případu s důrazným poukazem *relata refero* čili prodávám, jak jsem koupil. Musím ale dodat, že tato lvice byla kultivovaná paní, mluvila věcně a argumentovala rozumně.

Šlo o přijímací test adeptů studia o filmu; pro zachování anonymity se ani nepiďme, zda šlo o dějiny filmu na některé z filozofických fakult, anebo o FAMU ap. Součástí testu bylo i patnáct otázek k ekonomice kolem filmu. Adept měl rozhodnout, zda *rekupace* je navrácení prostředků, které do filmu vložil a) producent, b) distributor.

Slovo nenajdeme ani v Akademickém slovníku cizích slov, ani v příručkách *Co v slovnících nenajdete* nebo *Nová slova v češtině. Slovník neologizmů*, ani v Českém národním korpusu. Použitím vyhledávacích programů jsme zjistili, že slovo se nevyskytuje v žádném textu na síti. Za takové situace nezbývá než poohlédnout se po zdroji přejímky: vždyť předpona i přípona jsou z mezinárodního slovotvorného materiálu, jehož původ je latinský. Nejčastějším zdrojem dnešních přejímek je – i pro slova s latinskými stavebními prvky – angličtina. A právě v anglických slovnících jsme začali nalézat první stopy: sloveso *recoup* je romanismus (slovo původu románského) přejatý z francouzštiny, který je ve významu ‚odškodnit, kompenzovat' doložen s dvojím předmětem (odškodnit někoho za něco) už roku 1664, v bezpředmětovém užití až roku 1861. Anglické slovníky, včetně největších oxfordských, uvádějí ke slovesu minimum derivátů. Illustrated Oxford Dictionary (1998) má substantivum *recoupment* (‚náhrada, odškodnění, kompenzace') a adjektivum *recoupable* (přibližně ‚nahraditelný') ale nikoli *'recupation'[rikapejšn]*, která by byla předlohou české *rekupace*. Dokonce ani využití slovesa ve funkci podstatného jména, které je pro angličtinu typické, nezaznamenává v případě *recoup* žádný slovník. Českou (?) podobu kdosi k té anglické přitvořil, a to poněkud pitvorně, a přidělal nám starost s výslovností. Zatímco „rikap" bychom říkali stejně bez problémů jako „rimejk" *(remake)*, váháme nad výslovností slova, k němuž je těžké najít model. *Remake* by měl sloveso „rimejkovat" a verbální substantivum „rimejkování" (psali bychom *remakeování*?), žádnou

„rimejkaci" (ta je nám neústrojná, asi jako by byla „dabace" místo dabingu) – a jak tedy číst to slovo z titulku? „Rekupace", anebo „rikapace"? U tazatelky jsme zaznamenali výslovnost „rekupace", ale to poznamenávám jen pro úplnost.

Na internetu je vidět, že sloveso *recoup* prodělává v angličtině konjunkturu. Význam ‚navracení vložených financí, jejich získání zpět' – používaný o prostředcích **distributora** filmu – se objevuje i v souvislosti s gamblingem, v pojišťovnictví aj. Pozoruhodně si se slovem zahrává britská společnost pro recyklování plastových lahví *RECycling Of Used Plastic containers*, zkráceně RECOUP, a také výrobce doplňku stravy, antioxidačního vitamínu, který doplňuje funkci jater podporou detoxifikace, takže přispívá k obnově, zotavení těla. (Věřím i já Bohu, že i naši potomci budou témuž říkati po česku a obejdou se bez rekupace i bez rikapování.)

Otázka, zda uchazeč o studium zná výraz **úzce** odborný, je u přijímacího testu na vysoce výběrový obor sama o sobě regulérní. Méně pochopitelné je, že se k testování volí nikoli výrazy ze sféry „u filmu", ale ze sféry „za filmem" (K. Čapek: Jak se dělá film). Je skutečně otázka, zda bez znalosti vcelku banálního účetnického pojmu, který je vyjádřen slovem věru nebanálním, nemůže být přijat ke studiu budoucí Vláčil, Forman nebo Vávra – ale ptejme se také, kolik taková neznalost váží v celkovém skóre uchazeče (čekal bych, že to mnoho nebude). Co je však pochopitelné nejméně, je to, že k testování byl zvolen výraz, jehož současná i budoucí existence ve spisovné češtině je sporná: zatím nemáme doloženo, že by *rekupace* byla živým, užívaným a srozumitelným jazykovým prostředkem.

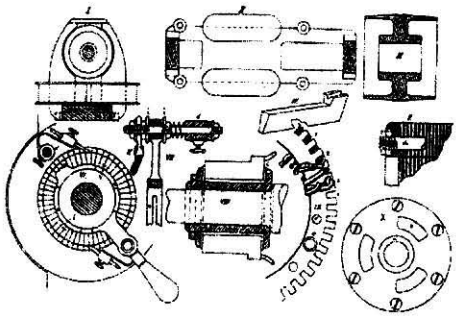
*Josef Šimandl*

# SLAVNÍ PŘEDCHŮDCI ČESKÉHO NÁRODNÍHO KORPUSU III

Vyobrazení základních periférií prvního elektrického serveru Havlíček, který umožňoval badatelům během několika minut prohledávat a statisticky analyzovat texty o délce až 1024 slov. Server Havlíček byl slavnostně uveden do provozu zástupci českých vlasteneckých organizací v podzemních prostorách budovy Akademie císaře Františka Josefa pro vědy, slovesnost a umění v květnu 1899, o dva roky později byl však po opakované důrazné výzvě vídeňské

Císařské a královské grantové agentury, z jejíchž prostředků byl pořízen, přejmenován na Prinz Eugen, der edle Ritter. Za zmínku stojí drobná, avšak nikoli nezajímavá skutečnost, že právě v souvislosti s mimořádnou úspěšností a spolehlivostí tohoto podzemního korpusu se v českých zemích postupem času vyvinula tradice umisťovat korpusová pracoviště pokud možno do suterénů a sklepů.

*Karel Kučera*



# Dialog mezi dvěma břehy

*Na budování Českého národního korpusu se podílejí jednak lingvisté, jednak matematici a programátoři. Jsou to dva rozdílné myšlenkové břehy, mezi nimiž dochází k výměně názorů, při níž se mohou něco nového dovědět obě strany. Tady je jedna z mnoha ukázek takových diskusí – mezi matematičkou RNDr. Jaroslavou Hlaváčovou (JH) a lingvistkou PhDr. Renatou Blatnou, CSc. (RB):*

**JH:** Občas se mi stává, že něco čtu a najednou nějaká věta jakoby vyskočí z textu a poleká mě. Přečtu si ji ještě jednou, a vida, tady je to! Jeden výraz, slovo, tu větu podle mého názoru kazí, nehodí se do ní. Nedávno jsem četla v novinách Metro článek o nejmodernější vojenské raketové lodi, kterou vyrobili v Rusku. V popisu lodi byla věta, která se mi nelíbila: „Jejím hlavním bojovým prostředkem je protiplavidlový raketový komplex Moskyt, který umožňuje během pouhých 35 vteřin vypálit salvu osmi raket, schopných zasáhnout *vytyčený* cíl na vzdálenost 10 až 120 kilometrů." Slovo *vytyčený* jsem zdůraznila já, a to proto, že mi v té větě nesedí. Já bych napsala *vytčený* cíl. Kdyby měl být *vytyčený*, musel by tam někdo

napřed jít a vyměřit ho. A to by pak bylo zbytečné se do něj z takové dálky trefovat, když už by tam někdo přímo na místě byl. Rozdíl pramení zřejmě z etymologie slova *vytyčený*. Pochází od slovesa *vytyčit*, což znamenalo původně *ohraničit* nebo *vymezit tyčí*. Teprve další význam znamená totéž co *vytknout*, tedy ‚určit, stanovit'. To jsem našla ve Slovníku spisovného jazyka českého.

**RB:** Ano, máš pravdu. Slovo *vytyčený* je odvozeno od slova *tyč*. Slovo *vytčený* souvisí se slovesem *tknout (se), týkat (se)* s významem ‚cpát, tlačit, tisknout něco někam'. Dokonavé *tknout* dnes už existuje pouze s předponami. Například: *dotknout se – dotýkat se*, staročeské *potknúti sě* ‚zmýlit se' – novočeské *potýkat se*, a také dvojice sloves, která nás zajímá: *vytknout – vytýkat*. Můžeme se ostatně podívat na přídavná jména *vytčený* a *vytyčený* do korpusu.

**JH:** Já už jsem se do korpusu dívala. Vyhledala jsem všechny tvary slov *vytyčený, vytčený* a pro úplnost i *vytýčený*. Jejich četnost jsem zapsala do téhle tabulky:

| | |
|---|---|
| vytčený | 168 |
| vytyčený | 182 |
| vytýčený | 29 |
| celkem | 379 |

Nejméně frekventované je slovo *vytýčený*. Nenašla jsem ho v žádném slovníku. Znamená to, že je chápáno jako nespisovné nebo nesprávné?

**RB:** Spíš nesprávné. Nespisovně by to znělo například „vytejčenej" a sama uznáš, že to zní dost divně. Já bych těch 29 výskytů připočítala k přídavnému jménu vytyčený. Máme tedy poměr 168 : 211.

**JH:** Tím se nám celá situace zjednodušila. Nechala jsem si v korpusu vypsat slova, která za nimi bezprostředně následují. Do tabulek jsem zapsala jenom ta, která se ve spojení s příslušným slovem vyskytují alespoň 3x.

*Vytyčený* (celkem 211 výskytů – 91 různých bezprostředně následujících slov):

| | |
|---|---|
| cíl | 38 |
| trasa | 16 |
| v | 9 |
| na | 8 |
| cesta | 8 |
| úkol | 7 |

směr 6
trať 6
záměr 4
hranice 4

*Vytčený* (celkem 168 výskytů – 58 různých bezprostředně následujících slov):

cíl 92
úkol 7
cesta 4
program 3

Moje intuice se tedy ukázala být správná, ale ne úplně. *Cíl* se skutečně pojí častěji se slovem *vytčený* než *vytyčený*, více než polovina „následníků" jsou *cíle*, ale počet *cílů* u druhého slova také není zanedbatelný. Další rozdíl mezi oběma slovy vyplývá z množství různých slov, která bezprostředně následují. Slovo *vytčený* jich má mnohem méně. Je to zřejmě tím, že se výrazně častěji pojí s již zmiňovaným slovem *cíl*. Některá slova se objevují v obou tabulkách. Na první pohled mě zaujal poměrně vysoký počet předložek *na* a *v*, které se vyskytují za výrazem *vytyčený*, ale nikoli za výrazem *vytčený*.

**RB:** Domnívám se, že s pouhým seznamem slov tady nevystačíme. Budeme se muset podívat na kontexty. Nejprve se podíváme na spojení *vytčený* a *vytyčený cíl*. Kontexty obou spojení jsou kupodivu velmi podobné: našla jsem jak *dosáhnout vytčeného cíle*, tak *vytyčeného cíle*, *cesta k vytčenému* i *vytyčenému cíli*. Převládá abstraktní význam. O konkrétním významu v souvislosti se slovem *tyč* lze mluvit jen u této věty: *„Hotovo" – družstvo se pustí a běží k vytyčenému cíli*. Docela mě zaujala tvoje zmínka o předložkách *na* a *v*. Přídavné jméno *vytyčený* je s předložkou *na* zachyceno 8x, ale jen tři kontexty mají konkrétní význam: *na trati vytyčené na silnici Strašice – Dobří*, *ochranné pásmo vytyčené na Olomoucku*, *soupeření v motocyklovém sprintu na vzdálenost čtvrt míle vytyčené na letištní ploše v Hoškovicích*. Předložka *v* nemá ve spojení se slovem *vytyčený* ani jeden konkrétní význam, jde o abstraktní užití převážně v politicko-ekonomických textech. Přečtu ti jeden příklad za všechny: *kritérium hospodářské konvergence vytyčené v Maastrichtské smlouvě*.

**JH:** Řekla bych, že *vytyčený* je obecnější. Vyplývá to jednak z již zmíněného počtu různých slov, která mohou bezprostředně následovat, jednak z jejich

charakteru. *Vytčený* se téměř nikdy nepojí s konkrétním podstatným jménem. *Železniční trať* nemůže být *vytčená*, pouze *vytyčená*. Slovo *cesta*, které se zde objevuje, je ve všech případech chápáno abstraktně (např. *vytčená cesta osudu*), nikdy jako konkrétní cesta (kamenitá, asfaltová...), po které se chodí.

**RB:** Vidím, že se docela dobře doplňujeme. Mohly bychom to shrnout tak, že abstraktní význam je dnes u obou přídavných jmen v psaných – a především publicistických – textech častější. Domnívám se ale, že naši předkové si skutečně své cíle, cesty, hranice a meze *vytyčovali*.

**JH:** Taky jsem ve Slovníku spisovného jazyka českého zjistila, že sloveso *vytknout* má ještě další významy. Ten nejčastější je ‚káravě upozornit na chybu, vadu'. Přídavné jméno slovesné vyjadřující tento význam se však objevuje jen ojediněle. Zde je zapotřebí připomenout, že u tohoto slovesa jsou dvě možnosti, jak tvořit přídavné jméno. Jednou je již zmíněné *vytčený*, druhou je *vytknutý*, které je však v ČNK zastoupeno pouze 2x. Naopak 2 výskyty výrazu *vytčený* se týkají káravého upozornění, výtky – *vytčené nedostatky*. Já bych napsala *vytknuté*. Co je správně?

**RB:** Samozřejmě *vytknuté*. A teď se zase zeptám já tebe: když vytknu v matematice nějaký výraz před závorku, je *vytknutý* nebo *vytčený*? Podle jednoho dokladu v korpusu *vytknutý*, ale ten nepochází z odborného či dokonce matematického textu. Takže co ty na to?

**JH:** Určitě *vytknutý*.

**RB:** Myslím, že teď už obě máme jasno. Já zas příště přijdu zkonzultovat nějakou statistiku, jo?

*Takových rozhovorů si tady užijeme hodně. Máte-li zájem, přijďte si taky popovídat, třeba na některý z našich seminářů. Informace o nich najdete na naší internetové stránce http://ucnk.ff.cuni.cz.*

*Renata Blatná a Jaroslava Hlaváčová*

Způsob řeči je obrazem ducha.
Seneca (O mravech)

# Já vůl, ty vole/ty jsi vůl, to je vůl...

Ne, to není začátek svérázného časování, ale jen naznačení mnohotvárnosti užití slova *vůl*, které tak často a ráda užívá v neformální a emocionální komunikaci především mužská část české populace. Že je to jeho dominantní využití, o tom nemusíme nijak pochybovat. *Český národní korpus*, tj. korpus dnešního psaného jazyka, nám dokládá, že takto se *vůl* užívá ve více než 80 % zaznamenaných výskytů; jeho frekvence v mluveném jazyce může být pochopitelně ještě vyšší. Jak se tedy vlastně užívá a jaká je jeho povaha? Ta otázka se může zdát vzhledem k tomu, že se užívá běžně, ať už aktivně nebo pasívně, někomu zbytečná. Ale je opravdu zbytečná? Stačí nám informace v SSČ, udávající vedle významu „dobytče" ještě to, že označuje zadruhé „hlupáka" s poznámkou v závorce, že jde i o nadávku? Nadávky jsou notoricky spolu s vulgarismy, některými expresívy apod. ve slovnících zvláště odbývané, snad proto, že je domácí uživatelé dobře znají a že (snad pro některé lidi) jsou neslušné a nestojí za to se jimi zabývat (nechme stranou nejasný postoj kodifikátorů v tomto případě, z SSČ není jasné, jestli jde o slovo spisovné nebo ne). Ovšem vyložit toto slovo jako „hlupák" a dodat, že jde o nadávku, říká velmi málo; slov pro „hlupáka" má čeština nespočítanou, ale rozhodně velkou řadu a slušelo by se mezi nimi rozlišovat. A co vše může být nadávkou (ve velmi různých smyslech), nikdo neví, dokud se do situace, kdy ji zažívá „v akci", nedostane. Hranice mezi urážlivostí a přijatelností je totiž obecně dána jen zhruba a je tu hodně individuálních možností a typů vnímání. Podívejme se na jazykový svět *vola* prizmatem ČNK.

Samostatné (syntaktické) užití slova *vůl* ilustrované prvními dvěma příklady v názvu, které je velmi časté, doplňuje ještě samostatné vokativní a pro některé, zvláště mladistvé mluvčí, téměř intimní oslovovací *vole*, které ovšem vůbec nadávkou není. Signalizuje jen formu kontaktu prostou (*vole* se užívá jak na začátku, tak na konci věty) nebo zesílenou a snahu zdůraznit zaujetí mluvčího (*ty vole!*).

Všechny příklady nahoře, které jsou ustálené, a jejich kontexty sémanticky ukazují, že označení adresáta těchto výrazů jako „hlupáka" spíš zavádí než něco kloudného vypovídá. Primárně jde o sdělení velké momentální nelibosti mluvčího vyvolané krajně nepříjemným opomenutím, selháním adresáta, popř.

o nesouhlas s ním. Kontexty nacházené např. ke zvolacímu výrazu *Je to vůl!* (a užívané v promluvě za ním) to dobře ukazují: ...*Zkomplikoval si život. Ztratí i ty brejle. Proč to neřek včas?* atd. Forma *Já vůl!*, zařaditelná spolu s mnoha dalšími do frazeologie, naznačuje, že tu mluvčí zřejmě nechce nutně považovat sama sebe za hlupáka v etymologickém, přímém významu (lidé sami sobě zase tak často nenadávají), ale jde mj. o výraz sebekritiky a lítosti nad zmařenou příležitostí. Jde tedy spíše o aspekty pragmatické než čistě sémantické. Všimněme si při této příležitosti i polysémie: vedle zesíleného kontaktového užití (viz výše) může *Ty vole!* ještě vyjadřovat i tento význam pocitu nelibosti a nesouhlasu (a podobně i *vy volové*). A co se s tímto slovem může dál v textu dít? *Vůl* je v této funkci (hlavně pro 3. osobu) zpravidla výrazně dál nemodifikovaný a stává v predikátu po kladném slovese *být* (jeho záporná podoba je řídká). Jediné standardní modifikace jsou běžně možné (obecně ovšem nijak často) v podobě intenzifikátorů. Z ustálených to jsou atributy *korunovanej, pěknej, úplnej, zatracenej, velkej, starej (vůl)*; neustálené jsou velmi řídké, srov. *senilní vůl*. K ustáleným modifikátorům patří i podoba užitá ve výrazu *Je to kus vola!*

Z výše položertovně nadhozené myšlenky konjugace (ovšem celé fráze, a ne samostatného substantiva) však může vyrůst otázka, jak je to pak s osobami dalšími, tj. v plurálu. ČNK dokládá úzus *vola* i v 1. osobě plurálu (zpravidla v podobě sebekritizujícího *my volové*), užívané jsou však i další osoby, i když pochopitelně s menší frekvencí. Časové zapojení, vedle prézentu, je možné pouze ve 3. osobě pro čas minulý (*Hospodář...byl pěknej vůl.*). Kondicionál je naproti tomu možný ve všech osobách (*To bys byl vůl, kdybys to připustil.*). Kromě nominativu a vokativu je jiné pádové využití tohoto slova možné jen neustáleně, ve volných syntaktických pozicích (*Povídejte tomu volovi něco o sobotě...*).

Kromě syntakticky samostatného užití (viz výše) nachází toto slovo běžné využití i uvnitř věty. Vedle nejobvyklejší pozice slova *vůl* v predikátu (zvláště po *být*, srov. ještě např. *Doktoři jsou volové.*) a jeho volné využitelnosti uvnitř věty (viz příklad výše a např. *Zabručel něco o volech...*) se občas objeví i další zapojení do role subjektu (*Ty volové tady po mně chtějí dvoudecáky...*) v neustálené promluvě. Zajímavá je i okrajová role atributová, srovnej např. *Momentálně chodí s volem Kočandrlem.*

Příručky slovo *vůl* slovnědruhově charakterizují jako nepochybné substantivum, o jiných možnostech mlčí. Je tu však situace opravdu tak jednoduchá? Známé a ustálené příklady jako *Ňákej/nějakej vůl (to sem dal)*, *Kerej vůl (volal tu pohotovost)?*, (*To umí/dělá) každej vůl* však jasně ukazují na pronominální

funkci (zájmeno neurčité a tázací, pochopitelně v daných kombinacích). Věta typu *Proč já vůl nezůstal radši doma!* pak pro svou ustálenost signalizuje spíše než prostou syntaktickou vsuvku běžnou víceslovnou partikuli. Konečně vokativní kontaktové *vole* má jasnou funkci interjekční.

K ustáleným formám s komponentem *vůl* patří (srov. však i řadu výrazů výše) i mnoho frazémů. K nejběžnějším patří frazémy jako *dělat někomu vola* a hlavně přirovnání, srov. *(to je) krve jako z vola, to je jako dát volovi jahodu.* Přirovnání typu *jako starej vůl jsem mu uvěřil/naletěl* ap. se přitom vzácně v naznačené sebekritice obracejí na samotného mluvčího. Příznačné jinak je, že jde často i o frazémy propoziční, větné, a to často přímo přísloví. Připomeňme si pár příkladů: *Nedělej ze mně vola!, Nebuď vůl!, Když se štěstí unaví, sedne si i na vola., Starýho vola těžko učit tahat., Ať děláš, co děláš, psovi ocas nenarovnáš a volovi rozum nedáš., Viděl někdo většího vola než (Karla)!, Dyž seš vůl, tak seš vůl!*

K poslednímu typu snad lze už řadit i ustálené burianovské *císař pán je vůl*, které ovšem motivuje i další výrazy neustálené *(Soudruh prezident je vůl.* a častý pouliční nápis po 21. 8. 1968 *Brežněv je vůl.*). Plynule tak přecházíme k úzu neustálenému, který se však – kromě své občasné invence – svou povahou od uvedených nijak zvlášť neliší. Srov. např. (poloustálené) *Osud je vůl.* či doklady o atraktivnosti tohoto slova pro nová přirovnání, z nichž některá se postupně ustalují: *je nastydlej jak vůl, je čilej jak uvázanej vůl* apod. Ojedinělé je hravé využití *vola* v předvolebním hesle *Nevolte voly!* Zdá se konečně, že jako aspoň zčásti ustálené je pro leckoho stále ještě kdysi proslulé pražské *Servít je vůl*, nápis před pár desítkami let pilně šířený po různých místech především studenty techniky, a žertovná tehdejší novotvorba v podobě pseudoodpovědi postiženého *Všichni jste volové. Servít!*

Propriální využití nachází *vůl* podle ČNK jen jako komponent v názvu hospod typu *U černýho vola.*

Málo je v češtině takových podivuhodně mnohostranných a komunikativně užitečných slov.

*František Čermák*

# Jazyková poradna, korpus a internet (3): Acquis

Jeden z tazatelů se nám svěřil, že když komunikoval s ministerstvem zahraničí a projevil neznalost slova *acquis* (kolikpak čtenářů nemá rozpaky, jak toto slovo vyslovit?), sklidil údiv a zaznamenal signály, že pro lidi z ministerstva není „in". Poměrně často se v médiích setkáváme se stesky nad tím, že naše státní a zastupitelské orgány nedělají dost pro náš vstup do Evropské unie. Případ slova *acquis* ukazuje oprávněnost stesků v rovině jazykové: že se o Evropské unii a našem vstupu do ní nepíše a nemluví natolik, aby se příslušné výrazivo stalo běžným; zároveň je tento případ dokladem jistých zdaleka ne jen jazykových pnutí v Unii.

Nízká informovanost širší veřejnosti není specifikum České republiky: průměrný Dán nebo obyvatel Kypru neví o Unii o mnoho víc než průměrný Čech. V německém sborníku „Evropská unie od A do Z. Příručka evropské integrace" (překlad vydalo Karolinum, Praha 1997) najdeme oddíl Abeceda evropské unie v heslech a ovšem také rejstřík, jenže pokud jde o *acquis*, kde nic tu nic. (Že by projev věkovité nechuti Němců ke galicismům? Že by projev řevnivosti, kdo má být první mezi rovnými čili kdo si má být rovnější?) Malá encyklopedie Evropské unie (vydal Ústav mezinárodních studií s přispěním PHARE, Praha 1997) uvádí hned jako své čtvrté heslo *acquis communautaire* s tímto vysvětlením:

> Odborný výraz pro právně a politicky dosažený stav integrace EU. K acquis communautaire patří nejen právo Společenství v užším smyslu, ale také podpůrné mezinárodní smluvní právo členských států. Zároveň se k němu počítají politická rozhodnutí v rámci společné zahraniční a bezpečnostní politiky (SZBP) a spolupráce v oblasti justice a vnitřních věcí a také závazky členských států ve vztahu k dohodám, které v budoucnu budou uzavřeny. Výrazu acquis communautaire ve smyslu „nabytých výsledků (bohatství) Společenství" se používá především při jednáních o vstupu nových států do EU jako souhrnného označení podmínek, které musejí uchazeči o přijetí při svém vstupu do EU převzít jako právní a politické závazky.

Hleďme, takže teprve při vstupu, ne už během jednání o vstupu? Inu, realita je jiná, než jak by se dalo dovozovat z jakýchsi osvětových příruček. Ostatně zalistujte v československém tisku z roku 1990: podle tehdejších prohlášení našich i zahraničních politiků jsme letos měli v EU být už několik let a každoroční nové průtahy a odklady snad neunikly ani chronickým zapomnětlivcům. Mimochodem, i k „listování" starým tiskem lze korpus využít (onoho pradávného roku 1990 se ovšem nedobereme).

Na internetu (pokud se proderete množstvím textů, které obsahují nikoli řetězec *acquis*, ale delší *acquisition*) se najde nezkrácené *acquis communautaire* převážně ve francouzských textech; výraz by se měl číst [a´ki komyno:´teR]. V anglických textech se objevuje anglizované „Community *acquis*" (*acquis* kurzívou jako přejatý výraz) nebo zkrácené *acquis*, např. *the opening of the analytical examination of the acquis with Bulgaria...* (Presidency Conclusions Cardiff, 15 and 16 June 1998). Jako francouzské slovo by se tedy acquis měl číst i v českém kontextu [a´ki].

ČNK (korpus SYN2000) nás na holičkách nenechává: *a/Acquis* má celkem 97 výskytů. Z toho 53x je použito úplné sousloví *acquis communautaire* a zcela ojediněle je vidět snahu počeštit aspoň adjektivum: *komunitární acquis*. Často se používá *acquis* samotný, popř. v doprovodu přívlastků, které tento výraz pomáhají aspoň asociativně kamsi zařadit: *harmonogram Národního programu pro přijetí acquis* (to je jakýsi pleonastický nominativ jmenovací), *přijetí a prosazování acquis, acquis EU/Unie, členové budou muset acquis aplikovat, adekvátní postoj k sociálnímu acquis, acquis v tomto sektoru*. Případný tisk slova acquis kurzívou vzal sice konverzemi do ČNK zasvé, ale zachováno je 14 výskytů výrazu „acquis (communautaire)" v uvozovkách, které signalizují, že pisatel výraz hodnotil jako neobvyklý.

Poněkud mlhavý slovotvorný význam (acquis communautaire ‚unijní výdobytek'?, jakási ‚společenstvová akvizice'? ‚spolčovací zisk'?) je překryt významem terminologickým, který jsme citovali výše. Jistá značkovost výrazu *acquis* by nebyla na závadu, ale překážkou šíření se může stát právě náležitá francouzská výslovnost: dnes, kdy je u nás znalost zdrojového jazyka spíš výjimečná. V českých textech se tedy nejspíš budeme nadále potkávat se *vstupem do EU*, někdy s *přistoupením k EU* – a občas bohužel i s *přístupem k EU* (méně kultivovaní pisatelé totiž nedoceňují, že tento obrat má tradiční význam nikoli ‚přistoupení', ale ‚způsob, jakým k EU přistupujeme, jak se k jejím záležitostem stavíme') anebo s *přístupem do EU* (to se v češtině zatím používá jen v obratech jako *(ne)mít přístup do EU* ‚mít tam zavřené/otevřené dveře'. O *přijetí do EU* se mluví čím dál méně; snad je v tom snaha posilovat vědomí o nutnosti naší vlastní aktivity.

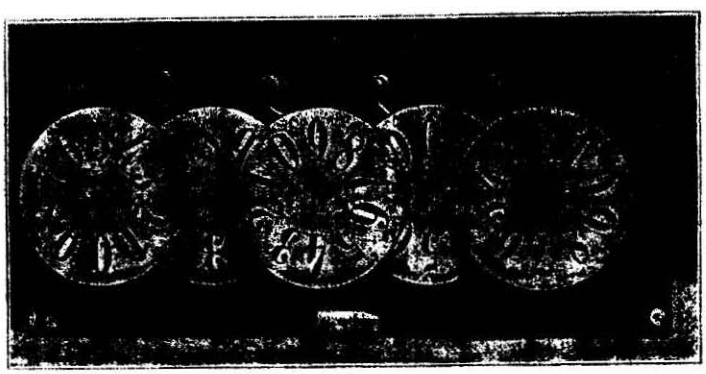
V jazykových prostředcích vyjadřujících onen kýžený vstup či přijetí tedy moc jasno nemáme, že? Ještěže je tu aspoň ten internet.

*Josef Šimandl*

# SLAVNÍ PŘEDCHŮDCI ČESKÉHO NÁRODNÍHO KORPUSU IV

Dobová ilustrace znázorňuje uživatelský interface modulu pro kvantitativní analýzu dat prvního paralelního česko-německého korpusu Libuše, budovaného od roku 1907 v Reichenbergu (nynější Liberec). Přes jeho poměrně malý rozsah a zanedbatelný podíl nářečního jazykového materiálu se korpusu Libuše dostalo – jako prvnímu a dosud jedinému korpusovému projektu u nás – oficiálního uznání císařského dvora. Důvodem ocenění byl živý zájem Františka Josefa I. o korpusovou myšlenku, rozdmychaný již na přelomu 80. a 90. let 19. století známým korpusovým extempore v Gebauerově studii *Über die Nationalität der Prager Gregoriusglossen*. K paradoxům korpusové historie patří, že dvorské intriky proti tzv. „korpusovému poblouznění" panovníkovu nedovolily císaři spontánně a přímo ocenit bouřlivě se rozvíjející českou korpusovou lingvistiku, a záminkou pro udělení rytířského kříže liberecké Libuši se proto nakonec musela stát zcela marginální skutečnost, že korpusu a jeho robustních nástrojů pro frekvenční analýzu se na počátku 1. světové války intenzívně využívalo k dešifrování nepřátelských vojenských depeší.

*Karel Kučera*



# POHLED Z BOHEMISTICKÝCH ŠKAMEN

Čemukoliv se učíš, učíš se pro sebe.
Petronius (Satiry)

## Kolísání rodu

Tématem tohoto článku jsou slova cizího původu, u kterých dochází ke kolísání rodu. Zvolili jsme tři skupiny slov podle zakončení a snažili jsme se s pomocí korpusu SYN2000 ukázat jejich tendence.

První skupinu tvoří slova zakončená na tvrdou nebo obojetnou souhlásku, která podle slovníků mají rod mužský nebo střední (a v druhém případě zůstavají nesklonná): *copyright, opus, depot [depo], foyer [foajé i foajér], bufet [bufet* i *bifé]*. V nominativu u slova *copyright* nemůžeme rozeznat rod, např. *měl i copyright ke každé fotografii*. V nepřímých pádech je slovo *copyright* sklonné, takže můžeme rod určit (v tom případě pouze mužský): *požaduje výhradní práva copyrightu, v souladu s copyrightem, výpověď smlouvy o copyrightech, výběr copyrightů, příspěvky za copyrighty*. V nominativu slova *opus* zaznamenáváme převahu tvarů mužského rodu: *Wagnerův životní opus, pětiminutový opus, zmíněný opus*. Střední rod můžeme zjistit především ve spojení *opus magnum*: *Joyceovo opus magnum*. V ostatních pádech má zmíněné slovo rod mužský, např. *Verdiho opusy*. U slova *depot [depo]* máme doložené jenom tvary mužského rodu: *uložen depot, velký depot, nový depot, hlubinné depoty, jeden z mála depotů*. U slova *foyer* je kolísání rodu výraznější než u předchozích příkladů, ale také vítězí tvary mužského rodu: *dvoupatrový foyer, starý i nový foyer, ve foyeru, do foyeru,* ale *po prázdném foyer, ve foyer činoherního studia*. Slovo *bufet* má ve všech případech mužský rod v souvislosti s šířící se tendencí k vyslovování toho slova podle způsobu psaní.

Na základě provedeného výzkumu můžeme usuzovat, že slova výše uvedené skupiny se čím dál tím více přizpůsobují českému skloňovacímu systému, a proto nesklonné tvary středního rodu jsou na ústupu ve prospěch tvarů mužských. Navíc u slov, u kterých výskyt dublet byl spjat s dvojí výslovností, zaznamenáváme sklon ke ztotožnění výslovnosti se způsobem psaní.

Další skupinu tvoří slova zakončená na *-e*, u kterých můžeme mluvit o kolísání mezi ženským a středním rodem: *aloe, agáve, ukulele, kánoe*. U slova *aloe* zaznamenáváme tvary ženské kromě několika případů, ve kterých není možné určit rod podle tvaru přídavného jména, zájmena nebo slovesa: *Proč je aloe léčivá?, šťáva z pouštní aloe Barbandensis Muller, aloe obsažená v... , mezi vzrostlými aloemi*. U slova *agáve* ve dvou případech na 14 výskytů můžeme zjistit ženský rod: *Žlutě kvetou americké agáve, takové agáve*, v ostatních případech rod zjistit nelze. Také u slova *ukulele* nejsme většinou schopni určit rod; tam, kde to lze, tvary rodu ženského a středního zaznamenáváme v podobném poměru: *malé ukulele, havajské ukulele*, oproti tomu *budou brnkat na maličkou ukulele*. Slovo *kánoe* má rod ženský až na několik výjimek: *Nejlevnější turistické kánoe stojí okolo osmi tisíc korun, každý z nás si pádluje své kánoe (?), spatřit turistické kánoe*.

Abychom shrnuli: u slov málo užívaných je těžké určit rod, u slov frekventovanějších je viditelná převaha tvarů rodu ženského.

Do poslední skupiny byla zařazena slova zakončená na měkkou souhlásku: *oranž, prestiž, trolej, esej, image [imidž]* – tedy slova, která mohou mít rod ženský nebo mužský. U slova *oranž* zaznamenáváme jenom tvary rodu ženského kromě několika málo případů, kde rod nelze zjistit: *sytá oranž, a hned za ním šla oranž a žluť, nalakovaným křiklavou oranží*.

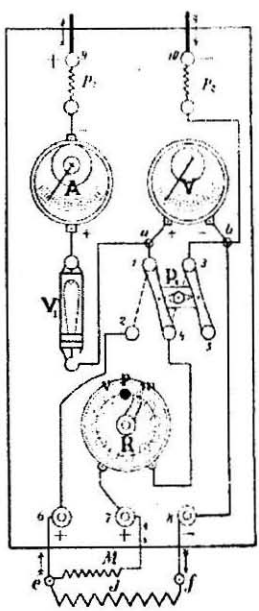
Podobná tendence je viditelná u slova *prestiž*: *Jelcinova prestiž, dějinná prestiž, léta budovaná prestiž, s mou nově nabytou prestiží, ze ztráty společenské prestiže*, a u slova *trolej*: *tramvajová trolej, pomocná trolej, nouzová provizorní trolej, s tramvajovou trolejí, do tramvajové troleje*. U slova *esej* zaznamenáváme tvary obou rodů s převahou tvarů mužských, např. *známý esej, ten esej, Jančarův esej, sto esejů, sbírka esejů*, a proti tomu *ironická esej, Camusova esej, knihu tvoří sedm esejí*. V některých případech můžeme mluvit o specializaci významu: když slovo *esej* označuje televizní pořad, má téměř ve sto procentech mužský rod: *tv esej, filmový esej, celovečerní autobiografický esej*. Podobnou tendenci zjišťujeme u slova *image [imidž]*, kolem 70 % tvarů má mužský rod: *autorův image, Klausův image, posílení feministického image*, kolem 30 % ženský rod: *podpořilo tu bondovskou agentskou image, specifická image* (ve spojení *self-image* jenom ženský rod), a také narážíme na ojedinělé případy použití tvarů středního rodu: *politik předvede voličům své image, bledé image jazzu, především mladé respondenty toto image výrazně oslovuje*. U slov zakončených na měkkou souhlásku vítězí v úzu jeden z dvou rodů, ale nejsme tady schopni určit žádné pravidlo.

*Urszula Lisowska*

# SLAVNÍ PŘEDCHŮDCI ČESKÉHO NÁRODNÍHO KORPUSU V

Schéma elektronického „mozku" přístroje na automatické dělení českých slov vyvinutého r. 1898 významným příznivcem a šiřitelem korpusové myšlenky Čechoameričanem R. Shimakem. Přístroj byl určen k mechanickému dělení slov na základě Shimakovy revoluční probabilitní gramatiky a pod obchodním označením *rozdružovač slov* se stal na dlouhou dobu komerčně nejúspěšnějším výstupem korpusové lingvistiky. (V zájmu objektivity dodáváme, že na závratném počtu prodaných dělicích zařízení se vedle vysoké kvality dělení slov podílela také skutečnost, že R. Shimak svěřil distribuci a prodej proslulému králi obchodních cestujících L. W. Robackovi, který rozdružovač českých slov s úspěchem prodával i analfabetům a cizincům, kteří se s češtinou nikdy nesetkali. Je známo, že mezi jeho stálé zákazníky, pravidelně si objednávající nové a nové verze rozdružovače, patřili Indiáni kmene Hopi, kteří nepřikládali valný význam zlatým valounům, hojně se nacházejícím na jejich lovištích, a pod vlivem Robackova obchodního hesla *Jediné, o co skutečně vždycky šlo, jde a půjde v lidském životě, je dělení slov!* se je snažili směňovat za zboží trvalejší hodnoty.)

*Karel Kučera*



Následující ukázky studentských prací vznikly v rámci volitelného semináře *jazyková variabilita*. Studenti, lépe řečeno v tomto případě studentky v zimním semestru 1999/2000 zpracovávaly témata, která si samy zvolily. Pracovaly však ještě s předcházející verzí Českého národního korpusu, která obsahovala přes 80 % publicistiky, tj. nikoli s reprezentativním korpusem SYN2000. V porovnání se současným reprezentativním korpusem v předchozí verzi na jedné straně některé jazykové jevy zastoupeny nebyly (např. *zelená krev*), zatímco v korpusu SYN2000 již zastoupeny jsou, na druhé straně některé pikantní příklady z novin (např. *chodit na kadibudku*) v důsledku redukce publicistických textů z 80 % na 60 % vypadly. To si ostatně každý uživatel korpusu může ověřit sám...

*Renata Blatná, vedoucí semináře*

# Několik slov o slovech expresívních (aneb hledat v korpusu mrtvoly – to není zabitý)

Sledujeme-li různé televizní politické diskuse, nemůžeme si nevšimnout, že účastníci těchto pořadů často používají, někdy dokonce využívají, ve svém vyjadřování expresivitu. Oním využíváním myslíme např. to, že expresívní prostředky slouží k aktualizaci výpovědi, k vyjádření důrazu, k vyostření konfliktu apod. Expresivitu v mluvené publicistice nese především zvuková rovina – např. změna intonačního rámce. Někdy jsou tyto změny provázeny také expresivitou v lexikální nebo syntaktické oblasti.

Vybrala jsem několik příkladů především adherentní expresivity, které jsou velmi časté v televizní mluvené publicistice. Zajímaly nás tyto výrazy: *mrtvola* (především *politická mrtvola*), *zabitý* (např. *senát*). Pokusila jsem se zjistit v Českém národním korpusu – pracovala jsem pouze s publicistickými texty –, jak časté je zde jejich použití v expresívním významu. Z oblasti inherentní expresivity jsem vybrala zakončení *-aření*, které je v současné době hojně využíváno k tvorbě expresívních slov (např. *politikaření*).

MRTVOLA: 300 náhodných výskytů – většinou s neexpresívním významem

Expresívní užití:

- *politická mrtvola* – 13x
- *jen přes jejich (moji) mrtvolu* – 3x

- *jde přes mrtvoly* – 4x

Větné příklady:

*Spolu s klesající produktivitou jednotky* (tj. člověka) *klesala přímo úměrně i její cena. A* ***mrtvoly****, jak známo, nepracují vůbec.*

*Bylo na čase, aby někdo překročil páchnoucí* ***mrtvolu*** *věčně hýčkaného komunistického útlaku.*

*Ta* ***mrtvola****, cos zasadil loni v zahradě, začala už klíčit?* (v básnickém textu – překlad básně T. S. Eliota)

*Na ledě by mohla zůstat* ***mrtvola*** (= hráč hokeje).

ZABITÝ: 494 výskytů – většinou ***zabitá*** *dívka, žena, mladík*

Expresívní užití:

- *spala jako zabitá* – 5x
- *Tahle sezóna je zabitá.* – 2x

Větné příklady a nadpisy:

*Tomu se nedá říci nevydařený závod, tady je na místě slovo* ***zabitý****.*

*Ostrava – tam je to* ***zabitý****.*

*Dny* ***zabité*** *fotbalem.*

*Křižovatka U* ***Zabitého****.*

***Zabitý*** *večer.*

ZAKONČENÍ ***-ření***: 797 výskytů

Eliminovala jsem výrazy typu *vaření, rybaření* apod. Nalezli jsme tyto výrazy: *politikaření* (159), *pletichaření* (15), *funkcionaření* (7), *slovíčkaření* (7), *melouchaření* (6), *spolkaření* (6), *intrikaření* (5), *kšeftaření* (4), *kramaření* (4), *příštipkaření* (4), *rodinkaření* (4), *tajnůstkaření* (4), *vladaření* (3), *figurkaření* (3), *fangličkaření* (2), *hokynaření* (2), *úplatkaření, čimčaření, bludaření, soudničkaření, revolucionaření, hloučkaření, šafaření, pomníkaření, batůžkáření, ptáčkaření, klepkaření, loutkaření, kritikaření, troškaření, hračičkaření* (po 1 výskytu).

Je zajímavé, že většina výrazů byla použita v souvislosti s politikou nebo s nějakou institucí (i výrazy jako *hokynaření – politické h., šafaření – š. s akciemi investorů, příštipkaření – p. tvoří tvář instituce*).

Ukazuje se, že expresivita v psané publicistice bují spíše v oblasti slovotvorné, oblast sémantická a kontextová je charakteristická více pro publicistiku mluvenou.

*Vendula Flaišmanová*

# *Čert* a *ďábel* na materiálu ČNK

**Čert** je dnes pro nás především pohádkovou postavou (do pohádek ovšem přešel z náboženství), a to postavou velmi oblíbenou. Dokladem je jeho častý výskyt v různých výčtech. Např. loutková hra „Kašpárek, obr a čert", *na maškarním bálu se rojili kovbojové, drákulové a **čerti**; ve společnosti trollů a **čertů*** apod. Velmi často se vyskytuje v trojici *Mikuláš, anděl a **čert**.*

Čert v pohádkách vystupuje jako bytost zlá, která člověka svádí k špatnostem a pak ho za to trestá, a proto je nutné se ho obávat. Toto pojetí je zřejmé v následujících ukázkách: *jsou to darebáci vhodní jedině pro kriminál a **čerta**; Rafinovaně strašil dceru, že v počítači jsou **čerti**. Obával se totiž, aby mu svými neodbornými zásahy něco nevymazala.* Variací na přísloví ***čert nikdy nespí*** je tento úryvek: *Ani zastánci termické likvidace odpadů neuměli srozumitelně vysvětlit, že spalovna není **čert**, jehož je třeba bát se vždy a všude.* V podobném významu se ***čerta*** užívá i v konkrétnějších přirovnáních. Např. *balkánští **čerti**, převlečení do srbské, chorvatské či muslimské kůže, budou posláni na exkurzi do nebes; hrozba kopancem na protřelé aparátnické **čerty** nemohla nikdy platit* atd.

Dosud se **čert** hodně objevuje v zakleních, kde si ještě udržuje expresivitu. Např. *oba psali ozlomkrk a **čerta** se starali o publikování; prezident poslal Klausův nápad k **čertu**.*

Stále se **čert** vyskytuje v přirovnáních a příslovích, která v novém kontextu (nejčastěji hospodářském) působí ozvláštňujícím dojmem: *investoři se vyhýbají riziku jak **čert** kříži; Také mě berou **čerti**, když vidím muže, který má na sobě nákladný moderní oblek a z nohavic mu koukají zaprášené a sešlapané boty.* Případně mohou být aktualizovány: *Kde není zdroj vnitřní poptávky, ani **čert** nekupuje.*

Poslední větší skupinu, kde se lze s **čertem** setkat, jsou spojení obsahující částici *čertví*, která vyjadřuje nejistotu. *Propachtovali si pozemky a **čertví** čí vlastně jsou. Být básníkem je házení perel sviním a **čertví** proč to dělám.*

Skoro vůbec se **čerta** neužívá k označení člověka bujného, veselého, popř. divokého, zlého. Přirovnání a přísloví se vyskytuje mnohem méně než ve slovnících a ovládáme je většinou jen pasívně. Expresivita už skoro vyprchala a pro zdůraznění se užívá vycpávkového slůvka *vlastně*. Viz např. věta *Čertví čí vlastně jsou!*

**Ďábel** je nejčastějším synonymem slova **čert**, ale na rozdíl od něho je dosud pociťováno jako expresívní. Však také často vyjadřuje intenzitu (hlavně ve sportu). Např. *běhali po hřišti jako **ďáblové**, brazilský **ďábel** Romário* (fotbalista), *jel jako **ďábel**, **Ďábel** impresário, alkoholový **ďábel**.*

Nezřídka se k **čertu** přirovnává na základě podobných vlastností nebo vzhledu. Např. *V Čechách už delší dobu převažuje obraz HZDS a Mečiara jako **ďábla**. Některé domorodé kmeny a Číňani označovali bělochy jako bílé **ďábly**.*

Se spojeními typu **dílo ďáblovo** či **ďábel – náš věčný Pokušitel a Svůdce** se můžeme setkat v náboženské literatuře. Hojně je **ďábel** zastoupen v kontrastech. Např. *Žádný z vás není ani **ďábel**, ani hotový svatoušek. Hrdinou je romantický pistolník s duší **ďábla** i anděla* (film Desperado).

Objevují se variace na známou situaci, kdy člověk *upisuje duši **ďáblu***. Např. *Upsala bys duši ďáblovi, jen abys vypadala líp. Televizní reportér je ochoten prodat duši ďáblu za senzační úlovek.*

Zvláštní skupinu tvoří přirovnání obsahující **čerta** i **ďábla**, popř. některé další synonymum jako **belzebub**, **satan** apod., a vyjadřující tak marnost počínání. Např. *stejně bezvýchodné jako vymítat hitlerovského **ďábla** stalinským **Belzebubem**.* **Ďábel** je tedy oproti čertu mnohem expresívnější a s tím souvisí jeho častější užívání v přirovnáních.

*Hana Matysková*

# Slovo *střecha* a jeho významy

Z etymologie slova *střecha* vyplývá, že se jedná o slovo všeslovanské, které ovšem nemusí ve všech jazycích znamenat totéž. Může znamenat např. jen ‚okraj krytiny' (ruština, ukrajinština) nebo jen ‚část přečnívající přes stěnu, tedy kryjící zásep' (lužičtina, bulharština). Slovo ***střecha*** náleží k litevskému *striegiu*, *striegti* – ‚pokrývati'. Samo ***střecha*** vzniklo ze **streg-sa*. To, že šlo o všeslovanský výraz, dokazuje výskyt tohoto slova v různých podobách i v jiných slovanských jazycích: staroslověnsky *strěcha*, rusky *strecha*, ukrajinsky *stricha*, polsky *strzecha*, hornolužicky *trěcha*, bulharsky *strjacha*, srbochorvatsky *streha*, slovensky *strecha* (Etymologický slovník jazyka českého a slovenského. ČSAV, Praha 1957, s. 479.).

V současných slovnících – SSČ a SSJČ – jsou uváděny tři základní významy slova *střecha* v češtině:

1. ‚plošný kryt chránící shora zvenčí zbytek budovy'
2. ‚kryt chránící shora vnitřek něčeho vůbec'
3. ‚okraj klobouku'

V Jungmannově slovníku z 19. století se vyskytují ještě významy další, dnes již nepoužívané – ***střecha*** jako ‚útočiště' či ‚krov'. ***Střecha*** u houby a ***střecha*** jako ‚stříška pro déšť', tedy ‚deštník'. Bezesporu zajímavější než významy základní je používání slova ***střecha*** v přeneseném slova smyslu.

Slovo ***střecha*** není pro omezený počet svých významů tolik frekventované ve frazeologii, vyskytuje se celkem v devíti frazémech, které uvádí „Slovník české frazeologie a idiomatiky" i „Slovník spisovného jazyka českého":

| | |
|---|---|
| 1. **s t ř e c h a světa** | ‚Pamír' |
| 2. **otcovská s t ř e c h a** | ‚rodný dům' |
| 3. **žít s někým pod jednou s t ř e c h o u** | ‚žít společně' |
| 4. **být pod s t ř e c h o u** | ‚být chráněný' |
| 5. **dostat něco včas pod s t ř e c h u** | ‚úspěšně něco dokončit' |
| 6. **mít vlastní s t ř e c h u nad hlavou** | ‚mít útočiště, domov' |
| 7. **posadit někomu na s t ř e c h u červeného kohouta** | |
| 8. **zapálit někomu s t ř e c h u nad hlavou** | ‚zapálit někomu dům' |
| 9. **vzít někoho pod svou s t ř e c h u** | ‚ujmout se někoho' |

Těchto devět frazeologických spojení je poměrně hodně užíváno i v dnešní češtině. Kromě nich se vyskytují v „Příručním slovníku jazyka českého" také spojení dnes už nepoužívaná:

| | |
|---|---|
| **bydlet pod nízkou s t ř e c h o u** | ‚bydlet v prostém obydlí' |
| **jako když hrách na s t ř e c h u sype** | ‚plynně číst' |
| **chytala mu s t ř e c h a nad hlavou** | ‚byl ve zlé situaci' |
| **hledět zpod s t ř e c h y** | ‚hněvat se' |
| **groš na s t ř e c h u hodí a spadnou mu dva** | ‚dobře se mu daří' |
| **vyvést ho na s t ř e c h u** | ‚někoho oklamat' |

Při porovnání významů uvedených ve slovnících a dokladů obsažených v Českém národním korpusu (ČNK) můžeme najít některé zajímavosti, které svědčí o nových významech slova ***střecha*** a o významových posunech tohoto slova. Velmi často lze v ČNK najít slovo ***střecha*** ve významu ‚záštita' či ‚ochrana': *pod střechou Socialistické strany arabské obrody tam vládne Saddám Husajn, legislativní střecha*, např. *nad ocelářstvím*. Zajímavá jsou v ČNK také

slova odvozená od slova ***střecha***. Některá z nich nejsou ještě příliš frekventovaná, ale je možné, že v blízké budoucnosti se stanou běžnou součástí našeho slovníku.

1. ***bezpřístřešní, bezpřístřešný*** – *spolek poskytne bezpřístřešní hře provozovací možnosti, román jako biografie zproblematizovaného individua v epoše transcendentální bezpřístřešnosti*
2. ***mezistřeší, mezistřešní*** – *mezistřešní prostory vlaku*
3. ***podstřeší, podstřešní*** – *severní část domu, kterým až do podstřeší prochází rozměrná šachta, podstřešní byty*

Zajímavé je v tomto ohledu zejména slovo ***podstřeší*** či ***podstřešní***, které jako by postupně nahrazovalo slovo ***podkroví***.

*Petra Smetáková*

# Chodit do, chodit na

Ve své ministudii bych ráda poukázala na drobné rozdíly v užívání předložek **na** a **do** ve spojení se slovesy, nastal-li nějaký významový či frekvenční rozdíl vzhledem k minulosti, která předložka dnes převládá apod.

Jelikož by látka byla příliš rozsáhlá, zaměřila jsem se pouze na vazbu se slovesem *chodit* (popř. *jít*), tedy *chodit na/do + substantivum*. Řídila jsem se převážně příklady z korpusu, čímž samozřejmě nastala další mírná redukce materiálu, neboť nelze zachytit stále se proměňující a pulsující hovorový jazyk.

Nejprve jsem se pokusila rozdělit užívání obou předložek do tří (často prolínajících se) skupin.

Do první jsem zařadila případy, kdy se objevují obě předložková spojení rovnocenně a význam zůstává prakticky stejný. Zajímavé je, že se jedná převážně o příklady související se školstvím a vzděláváním.
Lze tedy říci:

| | | | |
|---|---|---|---|
| chodit | **do (základní, střední, umělecké...) školy** | i | **na školu** |
| | **do gymnasia** | i | **na gymnasium** |
| | **do kurzu, kroužku** | i | **na kurz, kroužek** |
| | **do měšťanky** | i | **na měšťanku** |

| | | |
|---|---|---|
| **do jedenáctiletky** | i | **na jedenáctiletku** |
| **do přednášek** | i | **na přednášky** |
| **do zkoušek** | i | **na zkoušky** |
| **do plesů, schůzí** | i | **na plesy, schůze** |

Obecně lze shrnout, že ačkoli lze ještě stále zaslechnout obě varianty, postupem času zaostává genitivní spojení ***chodit do*** za progresívnějším, pohodlnějším ***chodit na***. Důkazem mohou být výše uvedené výrazy *chodit do plesů, do schůzí, do přednášek*, které se sice dnes ještě používají, ale mám dojem, že pro nastupující generace patří již mezi zastaralé. Stejný případ je i tradiční *chodit do hodin, do houslí, do piana, do baletu, do šití* apod.

Druhá skupina obsahuje případy, kde lze rovněž užít obou předložek, ale obsah výpovědi se posune nebo úplně změní.
Několik př.:

| | | | |
|---|---|---|---|
| chodit | **do divadla (obecně)** | ale | **na divadlo (konkr. žánr, představení)** |
| opačně: | **do (konkr.) diskotéky** | ale | **na diskotéku (obecně)** |
| | **do družstva (sport.)** | ale | **na družstvo (např. bytové)** |
| | **do krámu (nakupovat)** | ale | **na krám (prodavač)** |
| | **do práce (každý den)** | ale | **na práci (např. na stavbu)** |
| | **do vesnice** | ale | **(jít) na vesnici (přestěhovat se)** |

Ve třetí, nejpočetnější skupině jsou případy, kdy je vhodná pouze jedna vazba. Dalo by se říci, že předložka ***do*** se používá hlavně, jedná-li se o veřejná, společenská zařízení či objekty, zábavní podniky atd. Většinou jde o konkrétní budovu nebo uzavřený či alespoň ohraničený prostor, zatímco s předložkou ***na*** se pojí převážně abstrakta související s lidským konáním, zájmy, potřebami apod.

**Chodíme**

**do**: restaurace, hospody, kostela, chrámu, obchodu, supermarketu, nemocnice, kuchyně, kanceláře, kina, banky, muzea, klubu, knihovny, parku, laboratoře, dílny, sklepa, spíže, pokoje, posilovny, šatny, jídelny...

**na**: a/ <u>sport</u>: volejbal, fotbal, aerobic, hokej, gymnastiku, trénink, cvičení, krasobruslení...
b/ <u>koníčky</u>: houby, ryby, hon, jazyky (angličtinu, španělštinu...)
c/ <u>setkání</u>: jednání, demonstrace, svatbu, pohřeb, návštěvu, školení, konzultace, rodičovské sdružení, mše do kostela, úřady, poštu...

d/ medicína: akupunkturu, kontrolu, vyšetření, léčení, dialýzu, terapii, prohlídku, kosmetiku, masáže...

e/ jídlo: večeři, oběd, svačinu, pizzu, kuře, zabijačku...

f/ zábava: film, představení, koncert, muzikál, jazz, bigbít, ples, brigádu, dostihy, rande, flám, mejdan, zkušenou, kulečník, trh, hřiště, výstavu...

Ze sociologického hlediska jsme podle korpusu neobyčejně kulturní, vzdělaný, pracovitý a sportovně nadaný národ, neboť nejčastěji **chodíme do školy, do práce, do hospody, do posilovny a do divadla a na pivo, na aerobic, na koncert a na procházku.**

A na závěr jeden krásný příklad:

*Myslím, že občané a nejen obce Mrtník, pochopí, že je 1.5 m kanalizace nezachrání a s radostí budou ještě pár let* ***chodit na kadibudku*** *s optimistickou představou, že alespoň těmito ušetřenými prostředky, věnovanými pro rozvoj základního vzdělání, se jejich dětem dostane patřičného a pokud možno nezkráceného vzdělání, které je dovede k tomu, aby nezatíženou, tolerantní a vzdělanou myslí napravily chyby a „kocourkovské" nápady svých předchůdců.*

(Lidové noviny 1994)

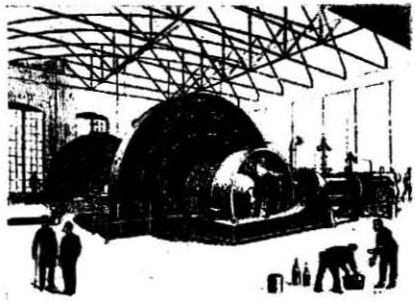
*Miroslava Kazdová*

# SLAVNÍ PŘEDCHŮDCI ČESKÉHO NÁRODNÍHO KORPUSU VI

Vzácně zachovalá dobová rytina zachycuje první pokusy s ukládáním jazykových dat na pevné disky (tzv. winchestery) při budování prvního českého diachronního korpusu *Svoboda*, které bylo zahájeno na podzim roku 1918. Na každý ze dvou vyobrazených ocelových disků o průměru 9,32 m bylo možno uložit až 2048 slov (včetně mezer a interpunkčních znamének). O tom, že česká korpusová lingvistika pracovala už v této době se skutečně špičkovou technikou, svědčí skutečnost, že v poválečné Evropě šlo o první dva pevné disky typu WOM instalované a uvedené do provozu severně od Alp a západně od Váhu. Technologie WOM (write-only memory) byla při výrobě pevných disků později opuštěna, třebaže podle řady navzájem nezávislých expertiz dosahovala při

ukládání dat dodnes nepřekonané spolehlivosti a přesnosti. K jejím nevýhodám patřilo, že po fixování do ocelového povrchu disku nebylo uložené informace možno už nijak získat zpět k dalšímu využití (podrobněji viz Smith, C.: *Disadvantages of one-way memory systems*. Lake Wobegon University Press 1997).

*Karel Kučera*



# Český národní korpus a jeho jazykové značkování

## 1. Úvod

Na konci října roku 2000 došlo na české lingvistické scéně k důležité události: byla ukončena první dlouhá etapa vývoje největšího českého lingvisticky zpracovaného zdroje jazykových dat – Českého národního korpusu (ČNK). Tento korpus počítačově uložených textů obsahující dnes již okolo 100 miliónů slovních tvarů má sloužit jako rozsáhlá reprezentativní databáze, které lze využít pro nejrůznější účely. Korpus je totiž zejména:

- základním východiskem pro budování nového velkého slovníku současné češtiny
- základem pro tvorbu nejrůznějších jazykových příruček, učebnic a vědeckopopulárních publikací o češtině
- studnicí lingvistických dat pro teoretické studium současné psané češtiny: pro studium její morfologie, syntaxe, sémantiky, pragmatiky a dalších úrovní a aspektů jejího jazykového popisu.

V tomto příspěvku se zaměřím na ČNK z hlediska morfologického a syntaktického a budu se zabývat jeho jazykovým značkováním, zejména pak značkováním morfologickým.

## 2. Jazykové značkování korpusu

*Jazykovým značkováním v širším smyslu* rozumím obecně přidání lingvistické informace k surovému nebo dosud jazykově nezpracovanému textu a/nebo formální, explicitní zachycení lingvistické informace, která je již v textu přítomna – každé jazykové značkování je tedy jistou lingvistickou interpretací textu, výrazem našeho pohledu na příslušný text. Toto značkování formálně vyjádřené tzv. *značkami (tagy)* má řadu podob podle toho, jaké informace chceme v korpusu mít a co chceme na jejich základě zkoumat. Základním typem značkování je segmentace surového textu, tj. rozčlenění nějakého textu, jehož

celistvá podoba se nazývá *dokument*, na přirozené úseky, jako je slovo, věta, odstavec, kapitola, část atd. Především členění textu do slov (slovem tu rozumíme i interpunkční znaménka, v tomto smyslu se také užívá pojmu *pozice v korpusu*) a vět je dáno jazykovými konvencemi a systémem jazyka. Na první pohled se zdá, že sotvaco může být jednodušší než rozčlenění textu na věty a slova. Máme-li např. na jednotlivá slova segmentovat větu

(1) Jeden z klíčových poslanců parlamentu, Jan Zahradil, se domnívá, že Evropská unie je pouze zbytnělou byrokratickou institucí.

nebude to ani nám, ani počítači činit velké problémy, neboť jednotlivá slova tu jsou jasně ohraničena a vymezena. Co si však počít s takovými slovy jako *oč, má-li, abychom, tys*? Slovo *oč* je stažená dvojice slov *o co*, spojení *má-li* je tvořeno dvěma zcela odlišnými částmi: *má* a *-li*, *abychom* je spojka vyjadřující 1. osobu čísla množného, ač spojky systémově osobu nevyjadřují, a *tys* je stažená dvojice slov *ty* a *jsi*. Jakým způsobem segmentovat takováto slova? Má smysl „pitvat" např. víceslovné názvy měst typu *České Budějovice*, kde obě dílčí slova, tj. *České* a *Budějovice* tvoří dohromady jediný celek? Není lepší takovýto název naopak chápat jako jeden nedělitelný celek a tak s ním pracovat? V češtině „bohužel" jednotlivé komponenty místních jmen nespojujeme spojovníkem, jak to činí alespoň v některých případech pravopis francouzský (srov. např. *Aix-en-Provence*) – bylo by to pro automatické zpracování jednodušší. A zase naopak: patrně není vhodné dělit např. slovo *československo-sovětský* na dvě části *československo* a *sovětský*. Obecně je pro další zpracování textu asi vhodnější pokládat ustálená spojení za jeden celek (nikoli však výše uvedené spojení typu *má-li*, které je naopak náležité rozdělit na dvě části). Vidíme tedy, že už na úrovni segmentace textu do slov mohou nastat problémy související pochopitelně s tím, jak chceme vstupní text dále zpracovávat.

Složitější problémy nastávají u segmentace textu do vět. Převážnou většinu vět lze segmentovat velmi snadno, jsou však věty, s nimiž jsou potíže. Uvažme např. větu

(2) Ke katastrofě židovského národa došlo roku 70 po Kr. Titus, syn císaře Vespasiana, tehdy dobyl Jeruzalém.

Člověk takový úsek „ručně" segmentuje snadno: okamžitě vidí, že jej tvoří dvě věty, z nichž první končí zkratkou *Kr.* včetně. Automatická počítačová segmentace je však v takovýchto případech velmi nesnadná. Jelikož druhá věta začíná vlastním jménem *Titus*, které začíná velkým písmenem, nemusí obecně řetězec *Kr.* znamenat konec věty, srov. např. větu

(3) Roku 70 po Kr. Titus, syn císaře Vespasiana, dobyl Jeruzalém.

Potíže tu působí víceznačnost tečky ukončující zkratku *Kr.*: tato tečka je totiž součástí zkratky (jejím posledním znakem) a zároveň může ukončovat větu zakončenou touto zkratkou. Pro automatické počítačové zpracování by bylo ovšem ideální, kdyby pravopisná pravidla nařizovala důsledné psaní obou teček, tj. *Kr.*., tak tomu však není. Tento problém ovšem nenastává u všech zkratek zakončených tečkou, nýbrž jen u těch, které se mohou objevit na konci věty. Je např. jasné, že v obsahově poměrně zajímavé větě

(4) Je třeba klást vysoké nároky na všechny pracovníky, tj. na všechny, kdo zajišťují výrobu.

neukončuje zkratka *tj.* ukončená tečkou celou větu, neboť tato zkratka se nemůže objevit na konci věty.

Uvedené příklady naznačují, v čem tkví hlavní problém větné segmentace: ve *víceznačnosti* jednotlivých řetězců a znaků a vůbec jazykových objektů. Tato vlastnost prostupuje celý jazykový systém. To samozřejmě není žádný objev, ale tento problém je jedním ze zásadních problémů počítačového zpracování přirozeného jazyka, jak uvidíme ještě jasněji, postoupíme-li dále, a to ke slovnědruhové a morfologické charakteristice jednotlivých slov textu. Celý problém je navíc komplikován velmi obtížnou uchopitelností lingvistických problémů – pro spletitou provázanost jazykových jevů nelze totiž metodologicky postupovat od jednoduššího ke složitějšímu. Chceme-li řešit nějaký problém na rovině $k$ jazykového popisu, potřebujeme využívat údajů na nižších ($< k$) rovinách popisu jazyka. Tyto údaje jsme však předtím získali na základě informací z roviny $k$, takže se točíme v kruhu. K řešení snazšího problému často potřebujeme vyřešit problém obtížnější, což je metodologicky absurdní. Např. správně segmentovat větu (2) můžeme teprve tehdy, máme-li k dispozici morfologickou analýzu textu (viz níže). Tuto analýzu však můžeme provést až poté, co jsme se vypořádali se segmentací. Právě vzájemná provázanost systému přirozeného jazyka představuje podstatný a specifický problém automatického zpracování přirozeného jazyka.

## 2. 1 Morfologická analýza

Poněvadž jazykový korpus slouží především výzkumu jazykovědnému, nestačí, aby obsahoval jen údaje o struktuře jednotlivých textů segmentovaných na úroveň jednotlivých vět, slov a interpunkčních znamének. Samozřejmě již údaje o samotných slovních tvarech, jejich konfiguracích a jejich syntaktické

spojitelnosti jsou velmi cenné. Nesmírné jazykové bohatství, jež korpus poskytuje, lze nicméně mnohem vydatněji vytěžit tím, že se jednotlivým slovním tvarům (dále jen slova) v textech korpusu přiřadí lingvistické informace z různých úrovní jazykovědného popisu. Další fází zpracování textu segmentovaného do slov a vět jsou tedy různé typy jazykovědných analýz, zejména *automatická morfologická analýza*, která každému slovnímu tvaru v korpusu automaticky (tedy počítačovým programem s pomocí slovníku) přiřadí jeho slovnědruhové a morfologické charakteristiky, a to v podobě tzv. *morfologických značek*; morfologickou analýzu zabezpečuje modul tzv. **morfologického značkování** – *morfologický analyzátor* (Hajič, 2000; Sedláček, 2000). Programový modul morfologické analýzy *(jazykové značkování v užším slova smyslu)* opatří každé slovo v textu automaticky lingvistickou informací dvojího typu:

- příslušným *základním tvarem*,
- veškerými jeho *potenciálními morfologickými interpretacemi*.

Tzv. **lemmatizací** je danému slovu přiřazena informace o jeho základním, slovníkovém tvaru zvaném *lemma*, popř. o více takových základních tvarech. K více základním tvarům patří

- slovo slovnědruhově víceznačné, tj. např. slovo *loučím*, které je buď tvarem 1. osoby jednotného čísla slovesa *loučit se* (*loučit se* je pak příslušné lemma), nebo dativ plurálu substantiva *louč* (to je příslušné lemma);
- slovo, které náleží více základním tvarům v rámci jednoho slovního druhu, jež jsou lexikálně víceznačné: např. sloveso *cenit-1* [na někoho zuby], *cenit-2* [o ceně – cenit si něčeho]. Lemmata jsou v dosavadní verzi značkování ČNK pojata poněkud šíře: některé lexikální jednotky jsou sloučeny do jednoho lemmatu (např. přechýlené podoby podstatných jmen mají lemma základního podstatného jména, od něhož jsou odvozeny; záporná podoba slovesa (např. *nekradu*) je zahrnuta pod lemma jeho kladného protějšku (tj. *krást*) apod.

Mimo lemmata přiřazuje morfologický analyzátor každému slovu všechny jeho **potenciální morfologické interpretace**, tj. informace o slovnědruhové příslušnosti daného slova a o jeho morfologických vlastnostech – např. informace o rodu, čísle a pádu podstatných a přídavných jmen, zájmen a číslovek, o stupni přídavných jmen a příslovcí, o osobě, slovesném a jmenném rodě, čísle slovesných tvarů atd. **Morfologická interpretace** daného slova je formálně vyjádřena **morfologickou značkou** tvořenou maximálně 15 údaji, z nichž každý je reprezentován jedním znakem na dané pozici, přičemž význam jednotlivých

pozic je jednoznačně stanoven (např. 1. pozice obsahuje identifikaci *slovního druhu*, 2. pozice obsahuje *jemnější kategorizaci slovního druhu* vymezeného hodnotou na 1. pozici, 3. pozice identifikuje *jmenný rod*, 4. pozice *číslo*, 5. pozice *pád* atd.) Každá morfologická značka je tak tvořena patnácticí znaků – např. morfologická značka NNMS1-----A---- přiřazená např. slovu *chlapec* označuje: N = substantivum, N = obyčejné (ničím nespecifické) substantivum, M = maskulinum, S = singulár, 1 = nominativ, A = kladná, nikoli negativní podoba slova – tato značka vyjadřuje protiklad typu *víra/nevíra*. Je-li dané slovo morfologicky, popř. lexikálně či slovnědruhově víceznačné (např. výše zmíněný tvar *loučím* nebo tvar *prosí*, což je slovesný tvar 3. osoby čísla singuláru i plurálu slovesa *prosit*, je morfologicky charakterizován dvěma odlišnými značkami, které se liší hodnotou na pozici čísla), opatří je morfologický analyzátor tolika značkami, kolik má toto slovo lexikálních, slovnědruhových a morfologických významů (včetně příslušných lemmat).

## 2.2 Automatická morfologická disambiguace

Automatická morfologická analýza textu, která spočívá ve správném přiřazení veškerých možných na kontextu nezávislých morfologických interpretací každému slovnímu tvaru v textu, je úkolem pracným, ale relativně nenáročným a lze ji obecně řešit prakticky bezchybně. Jakkoli má význam sama o sobě, lze ji především chápat jako nutný předstupeň k řešení úkolu nesrovnatelně náročnějšího – nalézt náležitou slovnědruhovou a morfologickou interpretaci daného slovního tvaru.

V konkrétním textu má každé slovo takřka vždy jen jedinou morfologickou, lexikální či slovnědruhovou interpretaci a tuto informaci je třeba ve značkovaném textu stanovit procedurou zvanou **morfologická disambiguace/ desambiguace** (zjednoznačnění), tj. výběrem správné interpretace ze všech interpretací poskytnutých morfologickým analyzátorem. Mluvčí jazyka (v našem případě češtiny) nemá s „ruční", tj. neautomatickou disambiguací zpravidla velké potíže (ač existují samozřejmě případy, kdy je výběr správné morfologické a slovnědruhové interpretace obtížný i pro člověka). Snadno např. určí slovní druh a morfologické vlastnosti všech slov věty

(5) Ty ženy snědí tři slanečky.

tj. bez problémů stanoví tyto slovnědruhové a morfologické charakteristiky:

*ty*: tvar=*ty*, lemma=*ten*, slovní_druh=*zájmeno*, poddruh=*demonstrativum*, pád=*nominativ*, jmenný rod=*femininum*, číslo=*plurál*

*ženy*: tvar=*ženy*, lemma=*žena*, slovní_druh=*substantivum*, pád=*nominativ*, jmenný rod=*femininum*, číslo=*plurál*

*snědí*: tvar=*snědí*, lemma=*sníst*, slovní_druh=*sloveso*, osoba=*3*, čas=*prézens*, způsob=*indikativ*, číslo=*plurál*

*tři*: tvar=*tři*, lemma=*tři*, slovní_druh=*číslovka*, pád=*akuzativ*, číslo=*plurál*

*slanečky*: tvar=*slanečky*, lemma=*slaneček*, slovní_druh=*substantivum*, pád=*akuzativ*, jmenný rod=*maskulinum_životné*, číslo=*plurál*

Bez ohledu na kontext je
slovní tvar *ty* také osobní zájmeno – lemma *ty*
slovní tvar *ženy* je také gen. sg., ak. pl., vok. pl.
slovní tvar *snědí* je také nom. a vok. pl. adjektiva *snědý*
slovní tvar *tři* je také nom. a vok. číslovky *tři* a též imperativ slovesa *třít*.

Je ovšem jasné, že obrovské objemy dat, jaké zahrnují současné jazykové korpusy, ručně disambiguovat nelze nebo je to možné jen s nasazením velkého počtu pracovníků zvaných *anotátoři*, kteří obvykle nejsou k dispozici (ručně disambiguován byl např. brněnský korpus *DISAM*, který ovšem obsahuje pouze cca 1 mil. slovních tvarů). Chceme-li tedy mít náležitě morfologicky jednoznačně označkované milióny slov, abychom mohli na rozsáhlém textovém materiálu studovat např. syntaktickou stavbu české věty, musíme ke značkování použít jiných metod než klopotné ruční práce. Je tedy třeba texty značkovat automaticky – softwarovými nástroji.

Co nejpřesnější a nejméně chybová automatická disambiguace textů přirozeného jazyka je obecně jedním z hlavních úkolů dnešní matematické lingvistiky, a to pro každý jazyk. Jeho obtížnost je ovšem do značné míry dána tím, nakolik je daný jazyk sám svým systémem víceznačný a nakolik má být značkování podrobné. Čeština je jazyk s obrovskou mírou víceznačnosti jak morfologické, tak slovnědruhové; existuje v ní přes 1000 tříd víceznačnosti. Tyto třídy jsou buď systémové – takovou třídu tvoří např. genitiv a akuzativ životných maskulin, jinou zase nominativ, akuzativ a vokativ singuláru substantiv středního rodu skloňovaných podle vzoru *město*, další třídou je instrumentál singuláru maskulina/neutra adjektiv a dativ plurálu adjektiv –, nebo nesystémové: tyto třídy obsahují náhodné homonymní tvary – např. výše uvedené *loučím*. Morfologicky disambiguovat český text programovými nástroji je tak úkol mimořádně složitý. Je to úkol nesrovnatelně složitější než automatická lemmatizace a automatická morfologická analýza, neboť výběr správné značky

z více teoreticky možných závisí na kontextu, v němž se dané slovo v textu vyskytuje, přičemž tu mohou hrát roli nejen faktory syntaktické, ale i sémantické. Zkušenosti – a to v celosvětovém měřítku pro různé jazyky – ukazují, že je to opravdu problém mimořádně komplikovaný: pro žádný jazyk (snad jen s výjimkou brazilské portugalštiny – Bick, 1996 – a snad i angličtiny – Karlsson et al., 1994) nebyl dosud uspokojivě vyřešen (viz níže).

Jak tedy tento úkol řešit? Existují obecně tři základní metody automatické disambiguace textů:

- *stochastická (statistická, pravděpodobnostní)*
- *pravidly řízená*
- *kombinovaná (stochastická a pravidly řízená).*

### 2.2.1 Stochastická disambiguace

V současné podobě jsou texty v ČNK automaticky disambiguovány programovým vybavením koncipovaným na základě stochastického (pravděpodobnostního) modelu, který je založen především na pravděpodobnostech přechodu mezi jednotlivými značkami v morfologicky analyzovaném (tedy dosud nedisambiguovaném) textu. Princip tohoto typu disambiguace spočívá v tom, že nejprve lingvista „ručně", tj. správně, označkuje jisté množství textů (o rozsahu řádově několika set tisíc slov, což je ještě únosné), a tak vznikne tzv. *trénovací korpus*. Statisticky koncipovaný disambiguační program – tzv. *tagger* – se poté „naučí" toto správné značkování, tj. učiní si představu o pravděpodobnostech přechodu mezi jednotlivými značkami a jejich četnostech, kterou si uloží do svých vnitřních tabulek („natrénuje se"). Zjistí např., že v textech se často vyskytuje dvojice (adjektivum, substantivum), přičemž oba elementy se shodují v rodě, čísle a pádě. Program, který se takto naučil správně označkovaný text, poté aplikuje své syntaktické (zejména slovosledné) „znalosti" vyvozené z trénovacího korpusu na dosud nedisambiguovaný korpus a tento korpus s větší či menší úspěšností disambiguuje.

Nejlepší programy pro stochastickou disambiguaci korpusů angličtiny dosahují až 97% úspěšnosti, úspěšnost morfologické disambiguace českých textů stochastickou metodou je zhruba na úrovni 94 % (konkrétně: 94,35 %) – tj. zhruba každé 16. slovo je disambiguováno chybně (do uvedené úspěšnosti jsou ovšem zahrnuty i slovní tvary, které mají jednoznačnou interpretaci bez ohledu na kontext; kdybychom brali v úvahu jen skutečně víceznačné slovní tvary, byla by úspěšnost stochastické disambiguace podstatně nižší – jen okolo 75 %). Uvedený rozdíl vyplývá zejména z odlišných typologických vlastností češtiny

a angličtiny – jazyk s poměrně velmi pevným slovosledem, jakým je angličtina, se stochastickými metodami založenými na statistické distribuci četnosti posloupností značek značkuje samozřejmě mnohem úspěšněji. Na rozdíl od angličtiny, kde se typické posloupnosti značek dané pevným slovosledem vyskytují přesvědčivě často, tj. s velkou četností, není v češtině k dispozici z hlediska slovosledu mnoho záchytných bodů, neboť čeština má slovosled dost volný. To tedy znamená, že slovosledných posloupností, které obsahují *n* značek a mají přibližně stejnou četnost, je více, a příslušný výběr náležitých značek je tudíž méně jednoznačný. Např. ČNK se dosud disambiguoval pouze stochasticky, v současné době se pracuje na vývoji pravidly řízené disambiguace (viz níže).

Hlavní problém, na který narážejí všechny stochasticky koncipované disambiguační programy (*taggery*), tkví v nedostatku tzv. *trénovacích dat*. Syntaktická variabilita textů je tak obrovská, že stochastické taggery se prostě nemohou naučit všechny možné posloupnosti značek. Pokud je rozdíl v četnostech různých morfologických interpretací výrazný (např. slovo *se* se jakožto předložka vyskytuje pouze v 9 % případů a jakožto reflexivní částice či reflexivní zájmeno se vyskytuje ve zbývajícím 91 % případů), je disambiguace poměrně úspěšná. Pokud se však četnosti různých interpretací sobě blíží, dochází k výběru nesprávné interpretace mnohem častěji. Zvláštností stochastické disambiguace je také to, že stochastické taggery na jedné straně jasnozřivě „uhodnou" správnou variantu ve složitém kontextu a na straně druhé se dopustí hrubé chyby v kontextech, kde příslušný výběr správné morfologické interpretace je (relativně) jednoznačný. Příkladem takové hrubé chyby je např. výběr nominativní morfologické interpretace substantiva následujícího v textu korpusu bezprostředně po předložce, což ovšem jazykový systém češtiny až na některé předložky cizího původu zcela vylučuje.

Úspěšnost disambiguace českých textů stochastickými metodami je tedy zatím dost nevalná. Jakým způsobem tedy morfologickou disambiguaci zlepšit – a to pokud možno výrazně?

### 2.2.2 Pravidly řízená disambiguace

Na jaře roku 2000 se skupina počítačových lingvistů rozhodla slovnědruhovou a morfologickou disambiguaci radikálně zlepšit (srov. Oliva et al., 2000). Autoři (Karel Oliva, Milena Hnátková, Pavel Květoň a Vladimír Petkevič) zvolili řešení, jež spočívá ve formulaci zcela exaktních kontextových **syntaktických pravidel**, která u daného slova buď určí správnou morfologickou interpretaci, popř. více

správných interpretací (je-li slovo v daném kontextu vskutku morfologicky víceznačné), nebo alespoň vyloučí co nejvíce interpretací nesprávných.

Podstatou této metody je intuitivní formulace desítek či stovek syntaktických pravidel, která odrážejí syntaktické konfigurace češtiny dané jejím systémem. Chápeme-li texty korpusu jako parolové projevy jazykového systému češtiny, pak pravidla zachycují právě tento systém, a tedy i jeho projevy. Pro ilustraci uvedu příklad takového pravidla.

**Pravidlo 1 - Předložkový lokál**

*Jestliže nikde ve větě nepředchází (ne nutně bezprostředně) před slovem* sl *žádný výskyt některého ze slovních tvarů* na, o, po, při, v, *není slovní tvar* sl *v lokálu.*

Pravidlo bychom ovšem mohli ještě podstatně zpřesnit a uvažovat jen předložkové a navíc lokálové interpretace uvedených slovních tvarů (víme-li např. na základě jiných pravidel, že některé z uvedených slov není předložka – např. se zjistí, že **při** je substantivum; nebo např. zjistíme, že některá z uvedených předložek nemůže mít v daném kontextu lokálovou interpretaci).

Pravidlo 1 se s výhodou uplatní např. při disambiguaci věty

(6) *Této souvislosti* nepřikládám bez bližšího *vysvětlení větší* význam.

V této větě není žádné slovo v lokálu, přičemž potenciálně (tj. bez ohledu na kontext) mohou být v lokálu slovní tvary vyznačené kurzívou. Pravidlo tedy odstraní lokálové interpretace slovních tvarů zvýrazněných kurzívou.

**Pravidlo 2 - Zvláštní tvary osobních a vztažných zájmen po předložce**

*Nechť slovní tvar* sl *je vztažné nebo osobní zájmeno začínající písmenem* n *a nechť* sl *není jeden z tvarů* nás, nám, námi. *Pak slovní tvar nacházející se v dané větě bezprostředně před* sl *je předložka (označme ji* p*). Navíc průnik pádů vyžadovaných předložkou* p *a pádů (většinou jde o jediný pád) tvaru* sl *je neprázdný, tj. předložka* p *se pojí se zájmenem* sl *– tvoří s ním předložkovou frázi.*

Použití Pravidla 2 ilustruji větou

(7) Nemůžeš se přece *podle něho* řídit.

Slovní tvar *něho* je tu osobní zájmeno (jednoduché pravidlo vyloučí teoreticky možnou interpretaci tohoto slova jakožto vokativu lexému *něha*). Bezprostředně předcházející slovo *podle* je podle Pravidla 2 předložka a nikoli adverbium.

Tato předložka jednoznačně vyžaduje genitiv, a tak tvar *něho* je nutně v genitivu (a nikoli v akuzativu).

Jak vidíme, uvedená pravidla mají dvě části: část popisnou a část výkonnou. Popisná část charakterizuje podmínky užití pravidla – syntaktickou konfiguraci, tj. příslušný kontext. V části výkonné se definuje akce, která se má s dosud nedisambiguovanými daty provést. Např. v Pravidle 1 je popsán segment věty nacházející se slovosledně před určitým slovem *sl* – tj. segment neobsahující ani jednu z předložek, jež mohou vyžadovat jméno v lokálu; část výkonnou tvoří direktiva, jež stanoví, že u všech slov nacházejících se v dané větě před *sl* se odstraňuje lokálová interpretace, pochopitelně pokud slovo takovou interpretaci potenciálně má.

Ukazuje se, že takovýchto jednoduchých pravidel lze pro češtinu formulovat několik desítek. Jádrem množiny pravidel však jsou (převážně dost složitá) **pravidla pro shodu a valenci**. Je to proto, že struktura české věty (a speciálně jednoduché věty, tj. klauze), jejíž slovosled je poměrně volný, spočívá na třech opěrných bodech: bohaté morfologii, shodě a valenci.

Kongruenční systém češtiny je komplikovaný a velmi rozvětvený, výhodou ovšem je, že pravidla pro shodu až na zcela marginální výjimky (speciální kontext), jež je zapotřebí „ošetřit" zvlášť, platí spolehlivě. Obdobně jako u Pravidla 1 se i tady s výhodou užije negativně formulovaných pravidel, tj. pravidel, jež vyvozují závěry z nekongruence. Např. ve větě

(8) *Tento nový typ se snažili* uvést do provozu co nejrychleji.

nemůže být vzhledem k neshodě v čísle substantivum *typ* podmětem, a tedy nemůže být ani v nominativu, a totéž se týká jeho shodných přívlastků *tento* a *nový*. Obdobně ve větě:

(9) *Tyto záležitosti nelze* přejít mlčením.

musí být podmět tvořen infinitivem, neboť jen takový podmět připouští přísudek *nelze*, tj. nikoli např. nějaké jméno v plurálu. Poněvadž podmět je již obsazen infinitivem *přejít*, nemůže se v blízkém okolí (toto *blízké okolí* se však musí zcela exaktně specifikovat) slovesného tvaru *nelze* vyskytovat nějaký slovní tvar v nominativu, neboť jinak by musel být podmětem a v klauzi nemůže být více podmětů (je ovšem jasné, že tento úsudek je jistým zjednodušením, neboť existují slovesa, v jejichž valenčním rámci jsou dva nominativy, např. sponové *být*: *Slušný* [nom] *nelze být za všech okolností*. Struktury s takovými slovesy je tedy zapotřebí v pravidle vyloučit). Ve větě (9) tedy ani slovní tvar

*tyto,* ani tvar *záležitosti* nemůže být v nominativu.

Valence tvoří kostru každé klauze a toho je třeba v pravidlech pro automatickou disambiguaci náležitě využít. Taková pravidla se pochopitelně neobejdou bez rozsáhlého a podrobného valenčního slovníku koncipovaného jak z hlediska povrchové, tak z hlediska hloubkové syntaxe. Pravidla mohou s výhodou využít jak obligatornosti určitého doplnění např. nějakého slovesa, tak toho, že aktanty mohou rozvíjet dané sloveso nejvýše jednou. Např. ve větě:

(10) Musíš dobře *vyslechnout názory* všech zúčastněných.

nemůže být slovní tvar *názory* vzhledem k neshodě s přísudkem v osobě i čísle podmětem věty, ale ani instrumentálem v plurálu, neboť ve valenčním rámci slovesa *vyslechnout* je obligatorní akuzativ a jiný potenciální akuzativ, který by mohl zaplnit příslušnou valenční pozici tohoto slovesa, ve větě není.

Z uvedených příkladů je zjevné, že autoři opírají svá pravidla o syntaktický systém češtiny. Níže shrnuji hlavní metodologické zásady, na nichž celá koncepce pravidly řízené morfologické disambiguace spočívá:

- pravidla odrážejí syntaktický systém češtiny, její langue
- pravidla jsou formulována pozitivně i negativně, tj. buď u daného slova (skupiny slov) rovnou určí správnou interpretaci, nebo vyloučí interpretace nesprávné – převážně jde o postupnou redukci morfologických interpretací až na (v nejlepším případě) jedinou správnou
- co nejvíce pravidel musí být formulováno bezchybně, tj. se 100% spolehlivostí, a to i za cenu, že se vyloučí jen málo nesprávných interpretací (tj. kontextové podmínky použití pravidla jsou co nejpřísnější), což je na úkor zvýšení přesnosti; jde především o to, aby pravidlo nevyloučilo nějakou správnou interpretaci jako chybnou
- nelze-li syntaktickými pravidly totálně disambiguovat, ponechají se alternativy a ty se pak (bude-li to ovšem zapotřebí) rozhodnou (např.) sémantickými pravidly; v žádném případě se „nehádá"
- pravidla jsou sdružována do skupin podle své spolehlivosti, přičemž nejdříve se aplikují pravidla ze skupin s nejvyšší spolehlivostí
- pravidla jsou na sobě vzájemně zcela nezávislá a nezávislou spoluprací postupně vylučují nesprávné interpretace nebo stanovují správné; uvnitř jedné skupiny se aplikují cyklicky, dokud jejich účinnost neklesne na nulu (tj. dokud nenastane situace, kdy se všechna pravidla z dané skupiny aplikovala, ale nic neprovedla)

- důležitou složkou celé koncepce je zpracování frazémů, kolokací a idiomů
- pravidla využívají (popř. hodlají využívat) veškerých zdrojů syntaktických informací, zejména různých slovníků: např. slovníky kolokací a idiomů, valenční slovníky, slovníky homonym a homonymních tříd atd.

Obecně není metoda pravidly řízené disambiguace ve světě příliš rozšířená, a to patrně z těchto (ne vždy oprávněných) důvodů:

- vyvíjet syntaktická pravidla je velice obtížné – vyžaduje to od autorů hluboký vhled do syntaktické struktury věty
- syntaktických konfigurací je v jazyce příliš mnoho, a tak se předem rezignuje na to, že by se systém jazyka dal vůbec postihnout
- neumíme-li zpracovat složité souvětí, rezignujeme i na věty jednodušší
- pravidla jsou jazykově závislá
- jednoznačně se podaří určit jen málo slovních tvarů, zbytek se stejně bude muset „hádat"
- má se zato, že úspěšnost nebude vyšší než disambiguace stochastickými metodami
- nepodaří se udržet celý systém pravidel, kterých bude muset být velice mnoho (řádově nejméně stovky), plně konzistentní.

Na základě dosavadní práce na pravidly řízené disambiguaci vyslovuji optimistický názor, že je možné vyvinout syntaktická pravidla, která povedou k vynikající, tj. takřka bezchybné disambiguaci českých korpusových textů. Je ovšem jasné, že takových pravidel budou nejméně stovky.

### 2.2.3 Kombinovaná metoda disambiguace

Třetí možností, jak automaticky disambiguovat elektronické texty v jazykových korpusech, je kombinace obou výše popsaných metod. Zdá se, že pro flektivní jazyky, konkrétně pro češtinu, je vhodné volit sériově spřažené programové moduly tak, že modul pravidly řízené disambiguace je předřazen modulu stochastickému, tj. výstup pravidlového modulu je vstupem do modulu stochastického. Tato koncepce vychází z toho, že

- stochastický modul se občas dopustí hrubé chyby při disambiguaci syntakticky jednoduchých struktur a pravidlový modul právě tyto jednoduché struktury dokáže sám bezchybně disambiguovat, takže stochastický modul se jimi pak už nezabývá
- pravidlový modul provede částečnou disambiguaci, tj. odstraní některé zcela jistě nesprávné interpretace nabízené morfologickým analyzátorem, a tak zúží stochastickému modulu prostor pro rozhodování, což má za následek

vyšší úspěšnost nejen stochastického modulu, ale disambiguace jako celku.

Experimentálně se zatím podařilo ověřit, že koncepce spřažených modulů dává poněkud lepší výsledky než samotný stochastický modul (94,35 % oproti 94,65 %). Jelikož práce na vývoji pravidlového modulu není ještě zdaleka u konce, není v současnosti možné porovnat úspěšnost spřažené disambiguace s úspěšností samotného pravidlového modulu.

### 2.2.4 Význam disambiguace pro moderní českou lingvistiku

Obecně má automatická morfologická disambiguace význam praktický a teoretický. Jejím cílem je:

- morfologicky jednoznačně interpretovat výstup z morfologické analýzy – tj. označkovat texty v korpusech češtiny
- připravit podmínky pro syntaktickou analýzu češtiny, tj. provést mělkou syntaktickou analýzu
- podstatně prohloubit teoretické poznání syntaktické (a také sémantické) struktury české věty.

Podívejme se podrobněji na vztah automatické morfologické disambiguace a automatické syntaktické analýzy věty. Součástí velkoryseji pojaté disambiguace by mělo být předběžné zpracování vstupního textu pro vlastní syntaktickou analýzu. Ta by měla tedy pracovat s co nejvíce jednoznačným vstupem a navíc s předběžně zpracovanými daty, přičemž toto předzpracování by mělo spočívat zejména v:

- identifikaci víceslovných slovesných skupin (např. v identifikaci složeného kondicionálu, opisného futura apod.)
- identifikaci reflexív (tj. spojení reflexívní částice s příslušným slovesem, adjektivem či substantivem)
- identifikaci slovesných významů tvořených více slovy (např. v identifikaci takových slovesných významů jako *dát pokyn*, *mít za to* apod.)
- identifikaci kongruenčně spjatých slov (v jednodušších případech)
- rozčlenění souvětí na klauze (v jednodušších případech)
- identifikaci jednoduchých předložkových nominálních skupin
- identifikaci koordinovaných skupin (v jednodušších případech)
- identifikaci kolokací a idiomů
- identifikaci relativně snadno rozpoznatelných valenčních vztahů.

Majíc k dispozici takto předzpracovanou větu, měla by se syntaktická analýza zabývat jen tím, co jí podstatně přísluší, tj.:

- identifikací větných členů
- identifikací syntagmat na všech hierarchických úrovních větné struktury (zejména stanovením valenčních vztahů a volných doplnění)
- vytvořením částečné či úplné struktury věty v podobě nějaké algebraické struktury (typicky např. strom, příp. les jako množina stromů, popř. komplexní závislostní struktura koncízně propojující závislostní vztahy s koordinovanými konstrukcemi – viz např. Plátek et al., 1984; Petkevič, 1995).

Na rozdíl od morfologické disambiguace bývá ovšem syntaktická analýza závislá na zvolené syntaktické teorii, zatímco ve stanovení hodnot morfologických kategorií slov a jejich slovnědruhové příslušnosti panuje v drtivé většině případů shoda (jsou tu však hraniční případy – méně jasné bývá např. vymezení částic a jejich odlišení od příslovcí). Problém syntaktické analýzy tkví především ve vysoké míře syntaktické homonymie (odhlížíme-li zcela od sémantiky), a analýza je tedy velmi obtížná (velmi často není jasné, které větné členy tvoří syntagma – viz např. známý problém určení syntaktické platnosti předložkové skupiny, tj. určení větného členu, na němž tato skupina závisí) a výpočtově náročná. Právě morfologická disambiguace by měla podstatně snížit kombinatorickou explozi, kterou je syntaktická analýza tak zatížena.

*Vladimír Petkevič*

Literatura:

Bick, E.: Automatic Parsing of Portuguese. Proceedings of the Second Workshop on Computational Processing of Written Portuguese. Curitiba, 1996, s. 91-100.

Hajič, J.: Disambiguation of Rich Inflection (Computational Morphology of Czech), Vol. I. From Morphology to Tagging. Praha, 2001.

Karlsson, F., Voutilainen, A., Heikkilä, J., Anttila, A.: Constraint Grammar. A Language-Independent System for Parsing Unrestricted Text. Berlin–New York, 1995.

Oliva, K., Hnátková, M., Květoň, P., Petkevič, V.: The Linguistic Basis of a Rule-Based Tagger of Czech. In: Sojka P., Kopeček, I., Pala, K. (eds.): Proceedings of the Text, Speech and Dialogue Conference TSD 2000, Brno 2000. LNAI 1902, Berlin–Heidelberg, 2000, s. 3-8.

Petkevič, V.: A New Formal Specification of Underlying Structures. Theoretical Linguistics Vol. 21 (1995), No. 1, s. 7-61.

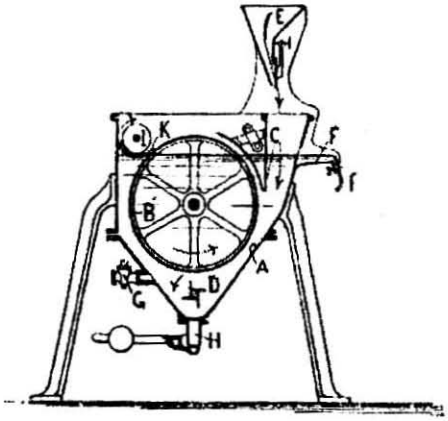
Plátek, M., Sgall, J., Sgall, P.: A Dependency Base for a Linguistic Description. In: Sgall P. (ed.): Contributions to Functional Syntax, Semantics and Language Comprehension. Praha, 1984, s. 63-97.

Sedláček, R.: AJKA - morfologický analyzátor češtiny. http://www.fi.muni.cz/~rsedlac/ajka.

# SLAVNÍ PŘEDCHŮDCI ČESKÉHO NÁRODNÍHO KORPUSU VII

Mezi neprávem zapomenuté milníky české korpusové historie patří skutečnost, že již v roce 1933 byl v rámci vytváření nástrojů pro plně elektrifikovaný korpus *Křižík* uveden do provozu první automatický slovnědruhový třidič. Na nákresu je zřetelně vidět jednak srdce celého zařízení, tvořené šestikomorovým systémem pro třídění slov do slovních druhů, jednak pravostranný vstup pro neurčené jednotky (input) a o něco níže znázorněný výstup pro jednotky slovnědruhově určené (output). Na třidiči z roku 1933 je dodnes oceňována přímočarost, s níž odrážel stav lingvistického poznání, resp. ochotu lingvistů shodnout se na slovnědruhové interpretaci řady pomezních případů. Přístroj vycházel ze zásady, že při třídění není třeba se řídit apriorními lingvistickými kritérii, vedoucími nevyhnutelně k řadě sporů, a třídil slova podle tzv. ekviprobabilitní gramatiky odvozené z metody *Monte Carlo*: do každé komory reprezentující slovní druh (počet šesti komor, a tedy i šesti slovních druhů byl zvolen z čistě konstrukčních důvodů) připadla vždy prakticky přesně šestina tříděných slovních jednotek a zdůvodnění, proč příslušné výskyty slov byly zařazeny právě do jednotlivých slovnědruhových komor, bylo ponecháno na následné lingvistické interpretaci. Koncepce třidiče realisticky počítala s tím, že schopný lingvista dokáže konkrétní výsledky ekviprobabilitního třídění přesvědčivě zdůvodnit, a účinnost tohoto přístupu se skutečně ukázala jako neobyčejně vysoká (pohybovala se mezi 99,893 a 99,998 %).

*Karel Kučera*



# Český národní korpus a lexikální archív Ústavu pro jazyk český v Praze

Nad prací našich předků se nám občas tají dech. Dívám se na fotografie důstojných pánů, kteří v roce 1906 pojali myšlenku, že kulturně jazyková situace češtiny dospěla tak daleko, že je nezbytné lexikograficky zpracovat moderní slovní zásobu. Celou věc inicioval lingvista František Pastrnek spolu s lingvisty Emilem Smetánkou a Josefem Zubatým. Ustavili v roce 1906 Lexikografickou a dialektologickou komisi III. třídy Akademie a začali přípravné práce na tezauru českého jazyka. Navštívili Mnichov, vyžádali si informace z Krakova, Bělehradu, Záhřebu a Petrohradu a poučili se o švédském slovníku.

Již v roce 1911 byla založena Kancelář Slovníku jazyka českého a začaly práce na budování archívu. Tehdy vznikla první excerpční pravidla. Excerpce zahrnovala literaturu od roku 1770. Excerpční lístek obsahoval v záhlaví slovo, kterého se výpis týkal, ve slovníkovém tvaru, byl uveden slovní druh, vazba nebo spojení, ve kterém se slovo vyskytovalo, kontext, dostatečně široký, aby z něho byl význam slova jasný, slovo, které bylo v záhlaví, bylo v kontextu podtržené, dole byl uveden zdroj, ve kterém bylo slovo nalezeno. O každém excerpovaném díle byl veden speciální záznam, tzv. zrcadlo, kde byly uvedeny všechny bibliografické údaje. Zpočátku byly excerpční lístky psané rukou, později na psacím stroji. Na excerpci se podílelo velké množství dobrovolných excerptorů, z velké části to byli středoškolští profesoři a učitelé, kteří pracovali v různých místech republiky. Naše řeč občas uváděla přehledy vykonané práce.

*Ukázka z Naší řeči z roku 1935*

*Z Kanceláře Slovníku jazyka českého*

Od 1. ledna do 31. prosince byly zpracovány tyto úkoly: 1. Spisy z r. 1834 (excerpoval Jaroslav Poch, profesor v Praze, lístků 3635). 2. Spisy Al. Jiráska (exc. Jarosl. Svoboda, úč. adjunkt fin. řed. v Bratislavě, lístků 10513). 3. Literární listy z 1881 až 1887 (exc. Jos. Soukup, uř. zemskeho úř. v Praze, lístků 3598). 4. Spisy K. V. Raise (exc. dr. Al. Získal, prof. v Praze, lístků 16000). 5. Naše doba (exc. Jos. Franc, ředitel reálky v Olomouci, lístků 13221). 6. Spisy Ter. Novákové (exc. Marie Holubová, prof. v Hradišťku u Kolína, lístků 3558). 7. Spisy J. Liera (exc. Jarosl. Malý, úř. knihovny Akad. domu v Praze, lístků 471). 8. Spisy Josefa Holečka (exc. Marie Ryšánková, choť profesora v Brně, lístků 7018). 9. V. Mrštík, Zumři (exc. Ant. Kácha, prof. v Praze, lístků 3560). 10. Spisy V. Štecha (exc. Žofie Vymetalová, prof.v.v. v Červ. Kostelci, lístků 5200). 11. Spisy B. Benešové, M. Majerové, M. Pujmanové, R. Svobodové (exc. Přem. Hájek, prof. v Praze, lístků 7396)

*atd., celkem dvacet šest položek. Úhrnem bylo od 1. ledna do 31. prosince 1934 odevzdáno 167.859 lístků.*

Excerpce se dělila na excerpci úplnou, částečnou a na tzv. pabčrkování, tj. vypisování zvláštností. Tzv. totální neboli úplná excerpce znamenala vypsání každého slova z daného díla jednou, ve všech významech, vazbách a frazeologických spojeních. Druhá excerpční pravidla vznikla v roce 1917. Byla zdokonalena na základě získaných zkušeností. Aby nebyla excerpována stále stejná slova a tvary, byla určena tzv. síta, díla, která byla vyexcerpována totálně, např. pro období 1854–1878 to byla Babička. Byl to vlastně soupis slov a významů již dostatečně doložených, která neměla být už dále vypisována Pro každý rok bylo excerpováno 10–15 knih různého obsahu. Zřetel byl brán především na významné autory. Do práce zasáhla rušivě první světová válka. Všichni zaměstnanci Kanceláře postupně odešli na frontu.

V roce 1935 se přistoupilo k vydávání Příručního slovníku jazyka českého, a to na základě asi 5,5 miliónu dokladů. K zpracování byl zařazen materiál od roku 1870. Ustoupilo se od nerealistického plánu zpracovat tezaurus. Paradoxně se první slovník jmenuje příruční, i když je ze všech tří akademických slovníků současné češtiny nejrozsáhlejší, má 9 svazků.

V Kanceláři Slovníku jazyka českého pracovali nejprve studenti, později středoškolští profesoři češtiny uvolnění pro práci na slovníku, všichni byli placeni z prostředků ministerstva školství. Podléhali tedy organizačnímu řádu škol, takže např. drželi i školní prázdniny. Za druhé světové války byl archív odstěhován do bezpečí do Mníšku pod Brdy. V roce 1946 vznikl na základech Kanceláře Slovníku jazyka českého Ústav pro jazyk český (dále ÚJČ). V roce 1952 byl začleněn do Akademie věd, která vznikla podle sovětského vzoru.

Ideologie v padesátých letech zasahovala i do výběru excerpovaných děl, totálně např. byli vyexcerpováni autoři Olbracht, Pujmanová, Majerová, Fučík. Po roce 1968 byla do excerpce zahrnuta díla všech funkčních stylů – nejen umělecká díla, ale i odborná literatura, publicistika a „projevy mluvené". Mluvený jazyk byl ovšem čerpán jen z krásné literatury, což ve skutečnosti není odpovídající zdroj pro studium mluvené formy jazyka. Intenzita excerpce ochabovala nejen vlivem různých politických okolností, ale i na základě chybných rozhodnutí ze strany odborného vedení.

Podle našich propočtů obsahuje archív ÚJČ:

1. a 2. materiál – excerpce od započetí do roku 1945, obsahuje asi 9 511 750 lístků.

3. materiál – 1945–1964 – obsahuje asi 343 620 lístků
4. materiál – 1960–1985 – obsahuje asi 128 790 lístků
5. materiál – 1975–1985 – obsahuje asi 128 790 lístků
6. materiál – od roku 1988 až do momentu, kdy se excerpce začala ukládat do počítače, – obsahuje 7 470 lístků

Podle těchto velice přibližných propočtů shromáždil lexikální archív ÚJČ asi za 80 let své excerpční činnosti přibližně 10 120 420 excerpt. Přehled nahoře jasně naznačuje, v jakém tempu se excerpovalo a jak postupně excerpce ochabovala. Excerpční lístky jsou uloženy ve velkých kovových a dřevěných kartotékách, kterými je obestavena chodba a pracovny prvního patra ÚJČ.

Na základě tohoto materiálu vznikly:

Příruční slovník jazyka českého (vycházel 1932–1957)
Slovník spisovného jazyka českého (1951–1970)
Slovník spisovné češtiny 1978, 1994 nové doplněné vydání (bylo doplněno na základě jiného, soukromého materiálu editora)

Již dlouho se na kartotékách i v jejich jednotlivých zásuvkách usazuje prach, do starého materiálu se přicházejí podívat ojedinělí zájemci.

Pokračovat v roce 1990 stejně jako dosud bylo nemyslitelné. Už šedesátá léta byla materiálově málo pokrytá. Množství vyexcerpovaných textů se zmenšovalo, až excerpce úplně ustala. Spoléhat na externí pomoc, tak jak tomu bylo v počátcích Kanceláře, už také nepřicházelo v úvahu. Změnila se společenská situace, lidé nebyli ochotní k dobrovolné spolupráci. Někteří odborníci se snažili hledat nové a současné řešení. Už v roce 1991 vznikl Počítačový fond češtiny. Zde se začali scházet zájemci, kteří chtěli koordinovat všechna pracoviště, která mohla pomoci tento ožehavý problém řešit. Tito lidé se také snažili nalézt pro nový projekt nutné finanční prostředky.

Jaké jsou přednosti excerpovaného materiálu?

Obsahuje jen cíleně vybraná slova, to znamená, že je v něm už obsažena analýza bezbřehého lexikálního materiálu. Necháme-li stranou závažný nedostatek ruční excerpce, a to malé množství získaných excerpt, pracnost a dlouhodobost jejich získávání, její skutečnou zásadní nevýhodou je pouze možnost vyhledávání podle abecedního řazení záhlaví. Takže např. není možné znovu zpracovávat jako celek vyexcerpované Máchovo dílo, protože je rozptýlené abecedně. Stejně tak není možné pracovat např. s jednotlivým funkčním stylem odděleně atd.

Díky počítačovému zpracování textů v rámci korpusu je velmi snadné odstranit všechny právě zmíněné nedostatky a nevýhody. Počet textů v korpusu neustále roste, jejich dostatečným množstvím bude odstraněno nebezpečí náhodnosti individuální volby při výběru zařazených děl. Lexikografovi i lingvistovi jsou nabídnuty všechny výskyty zadaného slova, které ve zkoumaných textech existují. Jak se zdálo na začátku práce na českém korpusu, množství výskytů, které korpus poskytuje, není „abundantní", tedy nadbytečné, jak se domnívala jedna zasloužilá lexikografka při zahájení budování korpusu, protože lexikografovi pomáhají při zpracování výskytů slova třídicí a statistické programy.

Podle anglického lingvisty Johna Sinclaira se lingvistika díky korpusům konečně dostává na úroveň přírodních věd. Před třiceti lety, kdy tento výzkum počítačového zpracování textů ve světě začal, bylo nepředstavitelné používat text s rozsahem o několika miliónech slov. Před dvaceti lety se o tom dalo uvažovat, ale hraničilo to se science fiction. Před deseti lety se o tom sice dalo přemýšlet zcela reálně, ale bylo to stále fantastické. Dnes je to myšlenka velmi populární.

Použitý materiál:

Červená, V.: Kancelář Slovníku jazyka českého. Naše řeč 1981/1.
Propagační příručka vydaná k 25. výročí založení Ústavu pro jazyk český

*Věra Schmiedtová*

Nic není věčné a jen máloco trvá dlouho.
Seneca (O útěše pro Polybia)

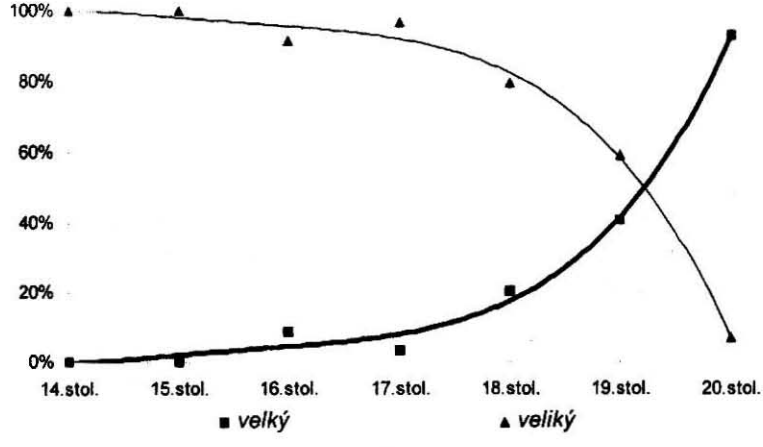
# Korpus a vývoj češtiny

Diachronní korpus, který vzniká jako jedna ze složek Českého národního korpusu, si klade za cíl vybudovat textovou základnu ke studiu vývoje českého jazyka. Třebaže období, které pokrývá (přibližně 7 století), mnohonásobně přesahuje časový záběr korpusu současné češtiny, soustředěného jen na několik posledních desetiletí, nemůže se co do velikosti ani zdaleka rovnat nedávno zveřejněnému stomiliónovému reprezentativnímu synchronnímu korpusu

SYN2000. Vzhledem k nutnosti manuálního přepisování většiny textů roste diachronní korpus jen pomalu (nikoli v řádu desítek miliónů slovních tvarů ročně, jak je obvyklé u jeho synchronního protějšku, ale v řádu desetitisíců, nejvýš statisíců) a jeho současný rozsah 1,8 miliónu slov o něm zatím rozhodně nedovoluje mluvit jako o korpusu reprezentativním.

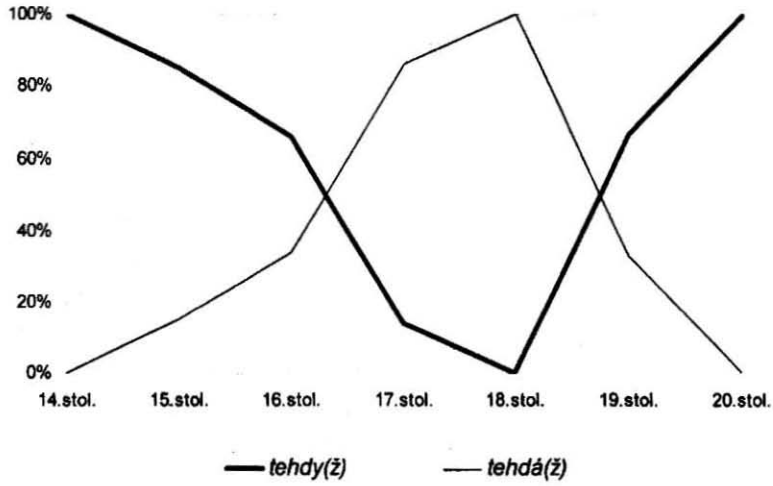
Bude muset uplynout ještě několik desítek let, než diachronní korpus bude s to spolehlivě zastoupit existující soukromé a institucionální soubory výpisků ze starších textů a poskytovat přijatelně přesné odpovědi na otázky toho typu, který odborníky i laiky zajímá nejčastěji (například: kdy se v českých textech poprvé objevil výraz *rajské jablíčko*; která slova jsou typická pro díla Jana Husa nebo Boženy Němcové a nevyskytují se u žádných jiných autorů ap.). Přestože jde o korpus, který je dosud spíše ve stadiu zrodu než ve fázi, kdy by s jeho využitím bylo možno navrhovat a ověřovat nové postupy zkoumání vývoje jazyka, je už dnes zřejmé, že v budoucnu bude s to jazykovědcům poskytnout víc než jen rychlou a podstatně univerzálnější náhradu za tradiční excerpci.

Jako ukázku takového obohacení našich dosavadních znalostí o historii češtiny lze uvést názorné a z korpusu snadno získatelné přehledy vývoje vzájemného poměru jednotlivých slov, slovních tvarů, podob ap. Graf A ukazuje, jak se v češtině měnil poměr mezi podobou *velký* (v grafu je znázorněna silnější čarou) a podobou *veliký*. Jak je zřejmé, ve starších obdobích vývoje jsou v českých textech doloženy téměř výhradně podoby s *-i-* (*veliký*), avšak postavení typu *velký* postupně sílí a ve 20. století už má naprostou převahu.




GRAF A

Trvalý vzestup jednoho prvku na úkor prvku jiného je však jen jedním typem vývoje poměru mezi dvěma slovními jednotkami nebo jejich podobami. Příkladem jiného druhu je dvojice *tehdy* (resp. *tehdyž*) a *tehdá* (resp. *tehdáž*). Z grafu B je zřejmé, že původně dominantní *tehdy(ž)* ustupovalo po čtyři století stále více do pozadí, avšak v novočeském období zažilo prudký vzestup vedoucí ke stavu, v němž je podoba *tehdy* opět dominantní a *tehdá* se uplatňuje jen okrajově.

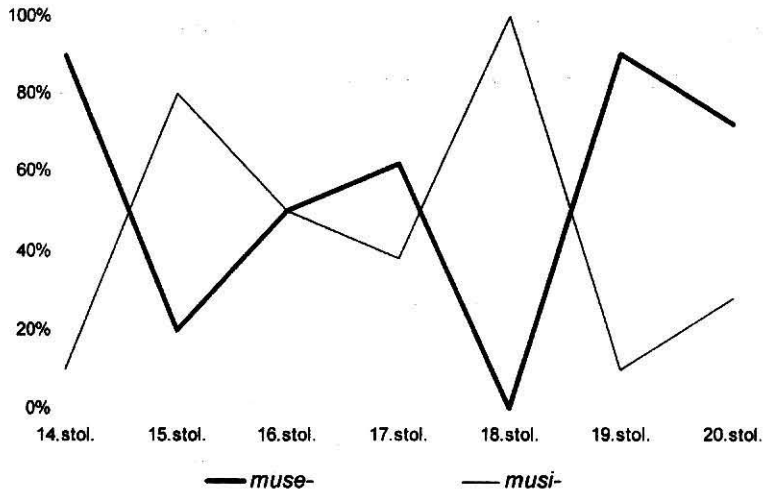



GRAF B

Jako poslední příklad – ukazující, že vývoj vzájemného poměru dvou jazykových jevů vůbec nemusí být tak jednoznačný jako v předchozích dvou případech – lze uvést sloveso *muset/musit*, respektive konkurenci jeho infinitivních tvarů 3. třídy (*muset, musel, musela...*) a 4. třídy (*musit, musil, musila...*). Graf C ukazuje, že obliba obou soustav tvarů během vývoje češtiny kolísala, aniž některá z nich nabyla skutečně dlouhodobé a jednoznačné převahy.

Průběhy křivek v uvedených grafech bude jistě nutno časem poněkud korigovat na základě širšího materiálu, ale není pravděpodobné, že by šlo o zásadní změny. Od využití rozsáhlejšího, reprezentativního diachronního korpusu lze navíc v budoucnu očekávat i podstatné rozšiřování a zdokonalování kvantitativních přístupů k vývoji jazyka, které nepochybně daleko překročí

rámec prostého srovnávání poměru jednotlivých slov a jejich podob, jaké jsme naznačili v uvedených příkladech.




GRAF C

*Karel Kučera*

Kdo začal, má hotovou půlku.
Horatius (Listy)

# CLARA: Corpus Linguae Arabicae a problémy počítačového zpracování arabštiny

Arabština představuje pro počítačové zpracování textů zvláštní problém. Tento problém tkví především ve způsobu záznamu arabštiny – arabské písmo (podobně jako některá další písma používaná pro semitské jazyky, např. hebrejštinu) totiž v naprosté většině nevyznačuje krátké samohlásky, i když pro jejich vyjádření existují příslušné symboly, a další značky užívané pro jednoznačné určení textu, jako značka pro zdvojení, pádové koncovky apod., se v textu objevují jen zřídka. Tyto značky se obvykle objevují v případě, že autor chce usnadnit čtení textu či chce zdůraznit některý gramatický jev. Plně vokalizovaný text (tj. text obsahující vyznačení samohlásek a všech dalších jevů, jako je zdvojení, neurčité členy, pádové koncovky apod.) se objevuje jen velmi zřídka, především v případě kanonických textů (Korán, Bible) nebo knih pro děti (slabikáře, pohádky pro nejmenší). Pro čtení textu je tedy třeba znát jazyk, a to jak jeho gramatiku, tak i významy jednotlivých slov, která se ve větě vyskytují.

Příklad:
Věta s plnou vokalizací:
*kataba 'l-mu'allifu riwájatan*[1]
napsal (ten) spisovatel román
(spisovatel napsal román)
Tatáž věta bez vokalizace:
*ktb 'lm'lf rwájt*

Je poměrně obtížné stanovit přesně poměr grafémů tzv. vokalizovaného textu (tj. textu s vyznačením samohlásek a dalších jevů, např. zdvojení souhlásky), protože arabština, jejíž spisovná forma je značně odlišná od tzv. arabských dialektů, kterými se hovoří v jednotlivých zemích, připouští ještě i tzv. substandardní normu, kterou je možno charakterizovat jako jakousi zjednodušenou formu jazyka. Tato norma se vyznačuje především tím, že i v mluvené formě často vypouští především pádové koncovky a neurčitý člen – toto vypouštění pak bývá nahrazováno pevnějším pořádkem slov, který jinak ve spisovné arabštině může být relativně volný. Nicméně při konfrontaci čistě

nevokalizovaného textu s textem plně vokalizovaným je možno konstatovat, že nevokalizovaný text představuje zhruba 63 % plně vokalizovaného textu. Jedná se však spíše o jakési vymezení hranic, protože jen málokdy v praxi narazíme na plně vokalizovaný text. Texty, které označujeme za nevokalizované, mohou obsahovat – a obvykle také obsahují – alespoň některé vokalizační značky.

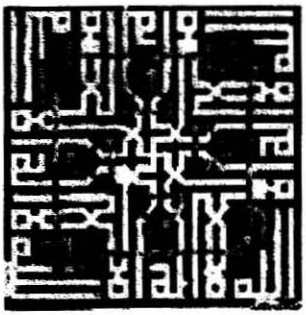
Už samotný tento fakt výrazně zvyšuje nejednoznačnost čtení jednotlivých slov. Například slovo *ktb*, použité v našem příkladu, lze číst například těmito způsoby:

*kataba* – on napsal
*kutiba* – on byl napsán (to bylo napsáno)
*kutubun* – knihy (mn. č.)
*kattaba* – nechat napsat

atd.

Tato nejednoznačnost (ambiguita) je dále prohloubena typem arabské morfologie. Ta je založena na tzv. konsonantním (souhláskovém) kořeni, který tvoří sémantický základ slova a je „aktualizován" pomocí samohláskových sestav a předpon a přípon. Např. kořen *drs* je spojen se studiem a vokalizované tvary *darasa* – „studovat", *darrasa* – „vyučovat", *darsun* – „lekce, přednáška", *dirásatun* – „studium", *madrasatun* – „škola" jsou příklady aktualizace tohoto kořene.

Kořenové souhlásky však mohou být pozměněny - například asimilací k předponě, nebo mohou úplně vymizet (velmi často u souhlásek *w* a *j*, ale i u dalších), což dále ztěžuje identifikaci kořene. Kořen je přitom tradičně považován za základní tvar a v podstatě všechny evropské (ale i staré arabské) slovníky jej používají jako položku, pod kterou řadí všechny ostatní odvozeniny. To znamená, že pokud chceme nalézt slovo *madrasatun*, je třeba zjistit, že se jedná o kořen *drs*, pod kterým ve slovníku najdeme příslušný tvar a jeho význam.



Podle posledních odhadů se v současných arabských textech používá zhruba 5000 kořenů[2] a přibližně 400 odvozovacích vzorů[3], z nichž většina je nejednoznačná. Na druhé straně neexistuje kořen, který by realizoval všechny možnosti, v průměru bývá od každého kořene realizováno kolem sedmnácti až osmnácti vzorů.[4]

Téměř každý tvar, založený na kořeni, je obvykle nejednoznačný. Jen málo vzorů je plně jednoznačných, u většiny tvarů existuje několik možností vokalizace (předpony a přípony jsou v textu vyznačeny, jedná se vždy o souhláskové morfémy). Počet těchto možností se obvykle pohybuje mezi dvěma až pěti, v některých extrémních případech však počet možností dosahuje pozoruhodně vysokého počtu. Například řetězec *j'd* může být interpretován jako tvar odvozený od několika kořenů:

- kořen *'dd* – „počítat"
- kořen *'wd* – „vrátit se"
- kořen *w'd* – „slíbit"
- kořen *'dw* – „běžet"

Vzhledem k tomu, že arabština umožňuje vytvářet pasívum pomocí vokalizace a jednotlivé tvary slovesa se mohou lišit jen koncovým vokálem, je celkový počet morfologických tvarů odvoditelných od tohoto řetězce značně vysoký – dosahuje dohromady 29 možností. Jde však o extrémní případ, kterých celkově není mnoho. Navíc i v těchto případech je z kontextu možné poznat, o které slovo se jedná.

Je však nutno konstatovat, že tato nejednoznačnost se projevuje především na úrovni jednotlivých slov. Už v případě tzv. bigramů, tj. bezprostředně následujících dvojic slov ve větě, je tato ambiguita značně menší. Právě tohoto rysu se běžně užívá pro další automatické zpracování textů.

Přes zmíněné problémy je takovýto text – po patřičné průpravě – pro člověka čitelný a srozumitelný; z hlediska počítače, který pracuje s těmito řetězci bez dalších informací (znalost významu, gramatiky, možných dalších vztahů), je čitelný a srozumitelný jen ve velmi omezené míře. Praxe ukázala, že softwarové prostředky, používané pro jiné jazyky, jsou pro arabštinu jen málo vhodné a že bude potřeba vyvinout programy specializované na arabštinu. Nejvíce pozornosti této problematice věnují firmy zaměřující se na strojový překlad z arabštiny, např. americká firma Apptec (http://www.apptec.com) nebo francouzská společnost CIMOS (http://www.cimos.com).

Přesto v nedávné době začaly být naplňovány předpoklady pro počítačové zpracování arabštiny. Rozvoj užívání počítačů v arabských zemích, rozšíření internetu a i další faktory způsobily, že elektronických, strojem čitelných arabských textů začalo rychle přibývat. Na několika místech se začalo pracovat – po vzoru elektronických korpusů angličtiny a dalších jazyků – na přípravě elektronického korpusu arabštiny. Jeden ucelený korpus, který je také plně gramaticky označkován, vznikl jako interní nástroj egyptské firmy Sakhr (rozsah 10 miliónů slov). Tento korpus je však přístupný pouze expertům této firmy a nezdá se, že by firma vůbec uvažovala o zpřístupnění tohoto korpusu dalším zájemcům. Další korpusy začaly vznikat relativně živelně, obvykle jako soubory textů, dostupných z internetu či jiných komerčních zdrojů. V dnešní době je relativně snadné shromáždit texty ve velkém množství – řada arabských periodik je dostupná na internetu, některé navíc prodávají své archívy, např. noviny al-Hayat vycházející v Londýně. Tyto korpusy jsou v naprosté většině dále netříděné, tzn. neumožňují například stylistické analýzy. Podle našich informací existují v současnosti dva korpusy arabštiny, které kromě těchto otevřených zdrojů získávaly texty pomocí vlastních sil, tj. především za pomoci načítání textů scannerem a následného rozpoznávání softwarovými prostředky (OCR programy). Jedná se o korpus textů na univerzitě v Nijmegen (Holandsko), připravený týmem pod vedením Jana Hooglanda (rozsah cca 11 miliónů slov, tříděný), a dále o elektronický korpus nazvaný CLARA (Corpus Linguae Arabicae), který byl zpracován na Karlově univerzitě v Praze[5].

Pražský korpus zpracovaný v Ústavu starého Předního východu a srovnávací jazykovědy UK FF je ze všech tříděných korpusů nejrozsáhlejší, v současnosti

dosahuje 43 miliónů slov. Jsou však předpoklady pro to, aby byl v blízké době rozšířen na 50 miliónů slov. Základním vzorem pro tento korpus byl British National Corpus (http://www.comp.lancs.ac.uk/computing/research/ucrel/bnc.html), vzhledem k lingvistickým a kulturním odlišnostem však byly provedeny některé změny ve struktuře korpusu, především v tematickém třídění textů. Korpus je sestaven z textů, které pocházejí především z 90. let[6], odráží tedy jazykový úzus poslední doby. Většina textů byla získána z otevřených zdrojů, nicméně zhruba 20 % textů bylo získáno pomocí scannování a následného načítání pomocí OCR programů. Korpus je zaměřen především jako základní referenční báze pro přípravu arabsko-českého a česko-arabského slovníku, z čehož vyplývá také skladba textů – poměrně velká část textů pochází z oblasti odborných publikací, např. korpus obsahuje výběr ústav arabských států, text několika zákoníků, značné množství ekonomických textů, ale i texty z oblasti techniky a přírodních věd. Beletristické texty tvoří zhruba 10 % celého korpusu a představují reprezentativní výběr z recentní egyptské a syrské literární produkce. Zhruba 50 % textů pochází z periodických tiskovin, kolem 35 % tvoří knihy, zbytek jsou příležitostné publikace.

Dalším problémem, který je u arabštiny nutno řešit, je – podobně jako u angličtiny – velké geografické rozšíření arabštiny a z toho vyplývající jazyková rozrůzněnost. V podstatě lze arabský svět rozdělit na dvě oblasti – tzv. arabský Východ (*Mašriq*) a arabský Západ (*Maghrib*). Druhá, západní oblast, je relativně silně ovlivněna francouzštinou, první oblast je považována za původnější z hlediska formy jazyka. Náš korpus se zaměřuje právě na arabský Východ, tj. především země Arabského poloostrova, Egypt a Sýrii. Podíl textů z ostatních zemí není příliš vysoký, celkově nepřesahuje 10 %.

Všechny výše uvedené informace jsou přirozeně obsaženy také v korpusu samotném, protože jsou důležité pro další analýzy užití různých forem arabštiny, syntaktických a morfologických jevů, lexikální zásoby atp. Kromě toho, kde

byly informace dostupné, jsou zařazeny také údaje o autorech (pohlaví, věk). Konečným cílem je vybudování tzv. vyváženého korpusu, v němž jednotlivé složky jazyka budou odpovídat také skutečné produkci v tomto jazyce, v našem případě ve spisovné arabštině. To je důležité především z důvodu spolehlivosti statistických údajů, které hrají značnou roli například při určování důležitosti jednotlivých významů a následně i při konstrukci slovníkového hesla.

V současnosti je tento korpus v tzv. syrové formě, tzn. k samotným textům není dodána žádná další informace. To je však z mnoha důvodů nevýhodné. Vzhledem k výše zmíněným problémům arabského písma se při vyhledávání například často stává, že vyhledávaný vzor zahrnuje jak jména, tak i slovesa, což přirozeně ztěžuje další analýzu dat excerpovaných z korpusu. Proto je žádoucí, aby korpus obsahoval i další, dodatečné lingvistické informace, především údaje o morfologických švech[7] a údaje o slovních druzích. To významně usnadní jak vyhledávání příslušných tvarů, tak i jejich další zpracování. Kromě toho tato forma korpusu umožní také další typy studií, například studium gramatických či syntaktických jevů.

Z těchto důvodů byly zpracovány dva další pomocné korpusy – jeden, v rozsahu 100 000 slov, má vyznačeny morfologické hranice, další (v současnosti cca 15 000 slov) obsahuje značky pro slovní druhy (spolu s určením dalších gramatických kategorií, jako např. rod, číslo atd.). Tyto dva korpusy budou sloužit pro vývoj nástrojů, které budou schopny zpracovat neznačený, syrový text a dodat do něj tyto značky.[8]

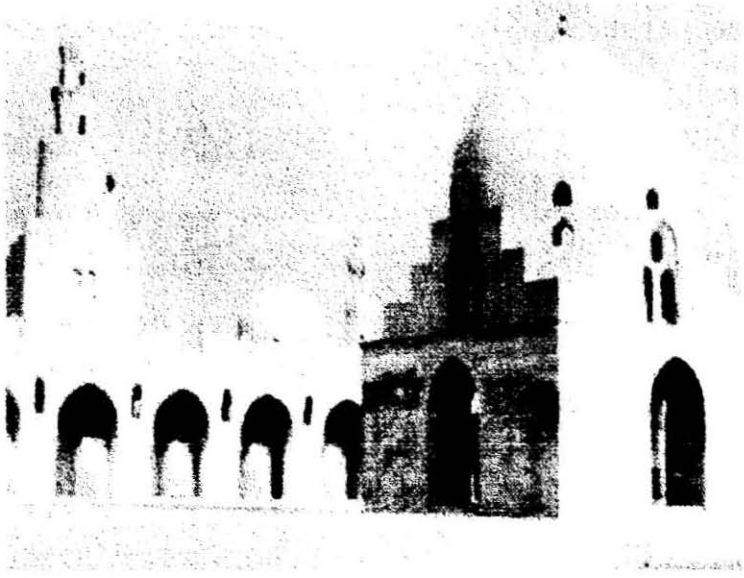
Všechny tyto nástroje jsou v současnosti přístupné pouze v Ústavu starého Předního východu a srovnávací jazykovědy Filozofické fakulty Univerzity Karlovy. Důvodem jsou především možné problémy s autorskými právy jednotlivých textů, ale i fakt, že projekt samotný je nadále zpracováván. Je však možno počítat s tím, že korpus bude zpřístupněn těm, kteří se na jeho rozvoji budou dále podílet, a to jak dodáváním dalších textů, tak i dalšími pracemi (spolupráce na morfologickém značkování korpusu, vývoj nástrojů pro tyto účely apod.). Již nyní se ukazuje, že o zapojení do tohoto projektu je značný zájem i v zahraničí, a v brzké budoucnosti se do jeho dalšího rozvoje zapojí několik zahraničních pracovišť. Takovéto spojení sil je možno jen přivítat, protože se bezpochyby promítne do urychlení dalších prací na vytváření korpusu.

Dodatek:

Některé internetové zdroje informací k tématu: Na internetu existuje celá řada serverů, které poskytují informace o arabských zemích, obsahují texty v arabštině apod. Zájemce o tyto servery

odkazujeme na arabské internetové portály, např. http://www.arabic2000.com, http://www.ayna.com, http://www.konouz.com a další. Zvláštní vyhledávač pro arabštinu vznikl jako speciální aplikace vyhledávace Altavista (http://www.arabvista.com). Dále uvádíme dva servery, které se přímo zabývají problematikou spojenou s tímto článkem, a internetovou adresu projektu CLARA: Xerox Arabic Morphological Analysis and Generation http://www.rxrc.xerox.com/research/mltt/arabic/ – online přístup, bez hesla, umožňuje zadání řetězce, provádí vyčerpávající morfologickou analýzu včetně významu slova/řetězce. Zpracoval Kenneth Beesley ve spolupráci s Timem Buckwalterem. QAMUS: Arabic Lexicography http://my.ispchannel.com/~tbuckwalter/ – obsahuje například seznam arabských kořenů, frekvenci arabských souhlásek v textech (zpracováno na korpusu arabských novin), informace o lexikografickém projektu Qamus („slovník"), přehled arabských kódových stránek a další informace. Zpracoval Tim Buckwalter. CLARA: Corpus Linguae Arabicae http://enlil.ff.cuni.cz/CLARA – základní informace o projektu, text o problémech počítačového zpracování arabštiny, základní odkazy na další relevantní zdroje. Zpracoval Petr Zemánek.

*Petr Zemánek*



[1] V tomto článku používáme zjednodušenou formu přepisu arabského písma do latinky tak, aby přepis co nejvíce odpovídal zvykům českého čtenáře. Současně volíme co nejjednodušší příklady.

[2] Beesley, K.R.: Finite-State Description of Arabic Morphology. In: Ubaydli, A. (ed.): Proceedings of the Second Cambridge Conference on Bilingual Computing in Arabic and English. Cambridge 1990, 7 stran, nečíslováno.

[3] McCarthy, J.: A Prosodic Theory of Non-Concatenative Morphology. In: Linguistic Inquiry 12, str. 373–418.

[4] Buckwalter, T.A.: Lexicographic Notation of Arabic Noun Pattern Morphemes and Their Inflectional Features. In: Ubaydli, A. (ed.): Proceedings of the Second Cambridge Conference on Bilingual Computing in Arabic and English. Cambridge 1990, 11 stran, nečíslováno.

[5] Tento korpus byl zpracován týmem ve složení: Petr Zemánek (vedoucí týmu), František Ondráš, Andrea Mustafá. V roce 2000 se členkou týmu stala také Naděžda Obadalová. Tento tým se zabývá především přípravou arabsko-českého slovníku. Projekt spolupracuje s pracovišti v Egyptě a Norsku – katedra arabského jazyka na univerzitě Ajn Šams v Káhiře a Ústav předovýchodních jazyků a kultur na univerzitě v Bergenu. V prvních třech letech (1997–1999) byl tento projekt podporován Grantovou agenturou České republiky a Grantovou agenturou Univerzity Karlovy.

[6] Časový limit pro zařazení textu do korpusu je rok 1975, toto pravidlo je porušeno pouze ve čtyřech případech – u Koránu, Bible, románu Zajnab od Muhammada Husajna Hajkala, který představoval po relativně dlouhou dobu jazykový vzor pro další autory, a u Džibrána Chalíla Džibrána, patrně nejslavnějšího arabského spisovatele.

[7] Řetězec v arabském písmu může obsahovat několik slov – jednak mohou ke slovu přistoupit různé částice, které se však v písmu píší dohromady, jednak arabština používá připojená přivlastňovací zájmena, která se také píší spolu s předcházejícím slovem. Například spojka „a" se v arabštině píše dohromady s následujícím slovem (wa-kataba – „a napsal"), stejně tak je tomu v případě předložky „v" (bi-), částice, určující budoucí čas (sa-jaktubu – „napíše, bude psát") apod. I v těchto případech jsou tvary relativně často homonymní, tzn. je možno je interpretovat několika způsoby a správná interpretace je možná teprve z kontextu.

[8] Systém značkování arabského textu je v současnosti odvozen od doporučení projektu EAGLES (Expert Advisory Group on Language Engineering Standards, http://www.ilc.pi.cnr.it/). Vzhledem k odlišnosti arabské morfologie však bylo třeba původní návrh značně upravit.

# KORPUSOVÝ KVÍZ

## Porovnejte své jazykové povědomí s korpusem

• Myslíte, že je možné příslovce *známě*? Dokázali byste ho použít ve větě?

• Máme slova *škrtidlo* a *zaškrcovadlo*. Snad se shodneme na tom, že vrah použije *škrtidlo*. Ale co když je zapotřebí omezit průtok žilami, hadičkou nebo finančními toky – používají se v těchto významech obě slova vyváženě, anebo má jedno z nich převahu?

• Jistěže *rolls-royce* je pěkný stroj. Má tvary podle vzoru *stroj*, nebo podle *hrad*, anebo se snad neskloňuje vůbec? Představte si: *nasedla do r...*, *jezdí v r...*

• Úprava dětských ponožek proti sklouznutí je *protiskluzná*, nebo *protiskluzová*?

• Jak byste nazvali takový ten trojkastrůlek spojený držákem, ve kterém se donášejí obědy?

• Dáte-li si modelovat nehty, bude to *modelování, modelace, modeling* nebo *modeláž nehtů*?

• Skloňujeme *Petrarca – Pertarky*. Platí to i pro jméno *Umberto Eco – Umberta Eka* atd.?

Svoje názory můžete se stavem v ČNK porovnat na následující straně – anebo přímo v ČNK.

## Porovnejte korpus se svým jazykovým povědomím

• Nalezli jsme *vypadat (důvěrně) známě*, *vypadal až bolestně známě*.

• *Škrtidlo* jsme našli 8x, z toho 3x obrazně (*rozpočtové š.*, *socialistické š.*), 3x jako vražedný nástroj, 1x jako nástroj zdravotnický (přitom nepočítáme výskyt

v textu o léčení v 18. století). *Zaškrcovadlo* máme doloženo 1x, a to v textu o první pomoci.

• Genitiv je doložen jen měkký (nebo snad nesklonný?) *do rolls-royce*. Lokál je většinou tvrdý: *v rolls-roycu* 7x, *v rolls-royci* 1x. Nemůže za ten nepoměr jedna jediná próza?

• 32x *protiskluzový*, 1x *protiskluzný*.

• Doložen je 2x *jídlonosič* (nepočítáme výskyt u Hrabala, kde slovo označuje osobu – mylně, místo *jídlonoš*) a 1x *obědník*.

• 20x *modeláž*, 1x *modelování*, 1x *modelace*. Bylo by ovšem namístě podívat se, nakolik je vysoký výskyt *modeláže* způsoben opakovanou reklamou v tisku.

• *Eca* 93x : *Eka* 33x; *Ecem* 17x : *Ekem* 4x; *Ecovi* 12x : *Ekovi* 0.

Tato zjištění učinila jazyková poradna ÚJČ AV ČR během posledních dvou let. Obsah korpusu se mezitím změnil. Nezapomínejme také, že ČNK nereprezentuje „správnost", ale obvyklost (v textech z 90. let). Přesto: Co o tom soudíte? Napíšete nám to?

*Josef Šimandl*

# Korpusový test

V roce 2000 byl dokončen zatím největší korpus češtiny SYN2000. Skládá se ze současných psaných textů převážně z let 1990–1999 a obsahuje 100 miliónů textových slov (z toho tvoří 15 % beletrie, 25 % odborná literatura a 60 % publicistika). Informace o jeho velikosti je jistě ohromující, ale co si pod tím představit? Co se z této záplavy slov a vět dá zjistit o naší mateřštině, kultuře a o tom, jak vlastně píšeme?

V následujícím příspěvku vám nabízíme třicet otázek, ve kterých bychom chtěli korpus a jeho materiál přiblížit formou testu. Můžete si tak sami ověřit, jaká je vaše jazyková intuice. Na konci najdete nejen správné odpovědi s širším komentářem, ale také vyhodnocení. Zvolili jsme formu testu, protože si myslíme, že si tak odpovědi lépe zapamatujete a navíc se i pobavíte. Neberte proto prosíme tento test ani jeho vyhodnocení jako zkoušku vašich znalostí, důraz jsme kladli

na komentované odpovědi. Většina otázek je založena především na frekvenčních údajích, v komentářích k odpovědím se však lze dozvědět i mnoho jiného o práci s korpusem a s jazykovými daty, která z něj získáme.

Každého občas napadnou otázky typu „Jaké je nejčastější vlastní jméno?" a odpověď pak nezbývá než odhadovat na základě vlastní zkušenosti. Ani my neznáme přesnou odpověď na tuto otázku, ale zato víme určitě, jaké je nejčastější vlastní jméno v korpusu SYN2000.

**Všechny následující otázky, i když to v nich není výslovně uvedeno, se týkají výhradně korpusu SYN2000.**

**1. Jaká je průměrná délka větného celku (tedy vět jednoduchých i souvětí – „od tečky do tečky") v korpusu SYN2000?**

a) 24 slov
b) 19 slov
c) 13 slov
d) 10,5 slova

**2. Kolik <u>různých</u> slovních tvarů najdeme v korpusu?**

a) 203 698
b) 526 041
c) 1 763 818
d) 5 426 322

**3. Vyberte nejfrekventovanější slovní tvar:**

a) a
b) v
c) se
d) je

**4. Kolik slovních tvarů se v korpusu vyskytuje pouze jednou?**

a) 145 278
b) 463 661
c) 821 223
d) 1 413 150

**5. Určete pořadí slovních druhů podle četnosti jejich výskytů:**

a) podstatná jména, přídavná jména, slovesa, zájmena, předložky, číslovky, spojky, příslovce, citoslovce, částice
b) slovesa, podstatná jména, spojky, předložky, přídavná jména, zájmena, číslovky, citoslovce, částice, příslovce

c) podstatná jména, slovesa, přídavná jména, předložky, zájmena, spojky, příslovce, číslovky, částice, citoslovce
d) předložky, spojky, podstatná jména, příslovce, přídavná jména, slovesa, zájmena, citoslovce, číslovky, částice

**6. Seřaďte interpunkční znaménka podle frekvence (od nejčastějšího k nejméně častému):**
a) tečka, čárka, otazník, středník, vykřičník
b) tečka, otazník, uvozovky, čárka, dvojtečka
c) čárka, středník, tečka, pomlčka, otazník
d) čárka, tečka, uvozovky, pomlčka, dvojtečka

**7. Určete pořadí nejčastějších předložek podle frekvence:**
a) před, pod, od, nad, k
b) v, na, s, z, o
c) o, v, s, z, k
d) na, za, v, k, u

**8. Určete pořadí nejvíce frekventovaných spojek od nejčastější k nejméně časté:**
a) a, že, i, ale, aby
b) a, ale, že, protože, nebo
c) a, i, když, kdyby, přestože
d) ale, jestliže, i, kdyby, že

**9. Které podstatné jméno se vyskytuje nejčastěji?**
a) člověk
b) den
c) rok
d) strana

**10. Které mužské jméno je nejvíce frekventované?**
a) Václav
b) Jiří
c) Petr
d) Jan

**11. Vyberte nejčastější ženské jméno:**
a) Marie
b) Hana
c) Jana
d) Eva

**12. O které zemi se v korpusu nejvíce píše?**

a) ČR
b) USA
c) Německo
d) Slovensko

**13. O kterém kontinentu se naopak píše nejméně?**

a) Antarktida
b) Austrálie
c) Afrika
d) Asie

**14. Které zahraniční město je jmenováno nejčastěji?**

a) Londýn
b) Moskva
c) Paříž
d) Bratislava

**15. O kterém zahraničním městě, které však není hlavním městem, se nejvíce píše?**

a) Mnichov
b) New York
c) Los Angeles
d) Jeruzalém

**16. Která země se nazývá „země orlů"?**

a) Slovensko
b) Albánie
c) Arménie
d) Rumunsko

**17. Které naše město se po Praze a Brně nejčastěji uvádí?**

a) Liberec
b) Ostrava
c) Plzeň
d) Olomouc

**18. Který ze dnů v týdnu je nejfrekventovanější?**

a) pondělí
b) pátek
c) sobota
d) neděle

**19. Který měsíc má nejvíce výskytů?**

a) listopad
b) leden
c) květen
d) září

**20. Které číslo je nejčastější (psané číslicí)?**

a) 0
b) 1
c) 2
d) 3

**21. Který rok je nejhojněji zastoupen? (připomínáme, že v korpusu převažují texty z let 1990–1999)**

a) 2000
b) 1994
c) 1995
d) 1996

**22. Která denní hodina je v korpusu nejčastější?**

a) 19.30
b) 20.00
c) 10.00
d) 18.00

**23. Jak jdou za sebou podle frekvence čtyři nejčastěji zmiňované politické strany?**

a) ČSSD, ODS, KDU, ODA
b) KSČM, ODS, ČSSD, KDU
c) ODS, ČSSD, ODA, KDU
d) ODS, KDU, ČSSD, ODA

**24. O kterých nositelích akademického titulu se nejvíce píše?**

a) MUDr.
b) Mgr.
c) JUDr.
d) Ing.

**25. Jak jdou za sebou z hlediska frekvence:**

a) dnes, včera, zítra
b) včera, dnes, zítra
c) zítra, dnes, včera

d) dnes, zítra, včera

**26. Seřaďte rozměry podle frekvence:**

a) výška, šířka, délka, hloubka
b) hloubka, výška, délka, šířka
c) délka, výška, šířka, hloubka
d) výška, délka, šířka, hloubka

**27. Seřaďte následující adjektiva od nejfrekventovanějšího k nejméně frekventovanému (pod jedním lemmatem (=základní slovníkový tvar) jsou započteny i všechny jeho stupně):**

a) vysoký, dlouhý, široký, hluboký
b) hluboký, vysoký, široký, dlouhý
c) dlouhý, vysoký, hluboký, široký
d) dlouhý, vysoký, široký, hluboký

**28. Které z následujících slov se v korpusu nejčastěji vyskytuje s pravopisnou chybou?**

a) tamější
b) vzpomněl
c) zapomněl
d) domněnka

**29. Určete počet výskytů chybného tvaru *dvěmi*. Počet výskytů správného tvaru *dvěma* je 15 882.**

a) 820
b) 116
c) 49
d) 0

**30. Který z názvů následujících jazyků je v korpusu nejčastěji zmiňován?**

a) polština
b) arabština
c) čínština
d) řečtina

**Vyhodnocení:**

| | A | B | C | D |
|---|---|---|---|---|
| 1. | 0 | 1 | 3 | 1 |
| 2. | 0 | 1 | 3 | 1 |
| 3. | 3 | 1 | 2 | 0 |
| 4. | 0 | 1 | 3 | 1 |
| 5. | 2 | 1 | 3 | 0 |
| 6. | 1 | 0 | 0 | 3 |
| 7. | 0 | 3 | 1 | 1 |
| 8. | 3 | 2 | 1 | 0 |
| 9. | 2 | 0 | 3 | 1 |
| 10. | 1 | 2 | 0 | 3 |
| 11. | 3 | 0 | 2 | 1 |
| 12. | 3 | 2 | 1 | 0 |
| 13. | 3 | 0 | 0 | 0 |
| 14. | 0 | 3 | 2 | 1 |
| 15. | 2 | 3 | 0 | 1 |
| 16. | 0 | 3 | 0 | 0 |
| 17. | 0 | 2 | 3 | 1 |
| 18. | 0 | 1 | 3 | 2 |
| 19. | 0 | 2 | 1 | 3 |
| 20. | 0 | 3 | 2 | 1 |
| 21. | 0 | 3 | 2 | 1 |
| 22. | 2 | 3 | 0 | 1 |
| 23. | 1 | 0 | 3 | 1 |
| 24. | 0 | 0 | 0 | 3 |
| 25. | 3 | 1 | 0 | 0 |
| 26. | 0 | 0 | 3 | 1 |
| 27. | 3 | 0 | 1 | 1 |
| 28. | 3 | 1 | 1 | 0 |
| 29. | 1 | 3 | 2 | 0 |
| 30. | 2 | 1 | 0 | 3 |

0 – 49 bodů

Vaše mínění a povědomí o korpusu ani představa o užívání jazyka nejsou nijak oslnivé a kromě toho máte zřejmě i smůlu, neboť pouhým náhodným

výběrem odpovědí dosáhnete výsledku okolo 40 bodů. A přece je tak jednoduché s tím něco udělat, podívejte se na naše internetové stránky a začněte si vyhledávat – jen tak pro potěšení a pobavení. Jazyk je přece krásný a stojí za to vědět o něm něco víc a z jiných úhlů pohledu – tak trochu jako ze zákulisí. Není to lákavé?

50 – 69 bodů

Vaše znalosti mají celkem solidní základy a v přemýšlení o jazyce nejste zdaleka nováček. Ovšem neškodí občas změnit úhel pohledu a odkryjí se vám nové obzory. Nebraňte se těmto novým pohledům a uvidíte, že vás překvapí a přinesou mnoho zajímavého a poučného do vašeho života.

70 – 90 bodů

Gratulujeme, váš výkon je skutečně vynikající. Máte jasnou představu o frekvencích slov v psaných textech a v korpusu. Máte velmi dobré předpoklady pro práci s korpusem a řadu zmíněných otázek byste si dokázali hravě vyhledat a vyhodnotit. Práce s korpusem pro vás navíc bude radostí a potěšením. Více než 80 bodů je až neuvěřitelný výsledek, musíte mít buď dobré špehy nebo špatné kamarády, kteří vám prozradili řešení předem.

**Odpovědi na jednotlivé otázky s komentáři:**

U některých frekvenčních údajů z korpusu SYN2000 uvádíme srovnání s frekvenčním slovníkem Jelínek, J. – Bečka, J.V. – Těšitelová, M.: Frekvence slov, slovních druhů a tvarů. Praha, 1961 (v textu dále značíme FSDT). Tento slovník zpracovává materiál v rozsahu 1 623 527 slov a v jeho složení je hojně zastoupena beletrie.

**1. Jaká je průměrná délka větného celku (tedy vět jednoduchých i souvětí – „od tečky do tečky") v korpusu SYN2000?**

Korpus SYN2000 obsahuje 7 639 336 větných celků, průměrná věta má tedy okolo 13 slov. Počet větných celků i jiné základní statistické údaje o korpusu je možné zjistit přímo v našem korpusovém manažeru GCQP zadáním volby menu Korpus > Souhrnné informace.

**2. Kolik různých slovních tvarů najdeme v korpusu?**

Korpus SYN2000 obsahuje 1 763 818 různých slovních tvarů.

**3. Vyberte nejfrekventovanější slovní tvar:**

Pro zajímavost uvádíme pořadí deseti nejfrekventovanějších slovních tvarů spolu s jejich frekvencemi v korpusu SYN2000:

| | |
|---|---|
| a | 2 690 157 |
| se | 1 997 092 |
| v | 1 836 848 |
| na | 1 532 219 |
| je | 950 984 |
| že | 858 341 |
| s | 711 281 |
| z | 654 966 |
| o | 618 915 |
| do | 589 261 |

**4. Kolik slovních tvarů se v korpusu vyskytuje pouze jednou?**

Korpus SYN2000 obsahuje 821 223 slovních tvarů, které se v něm vyskytují pouze jednou. Většinu těchto tvarů tvoří vlastní jména, číslovky a hlavně slova obsahující překlepy. Vzhledem k celkovému počtu všech různých slovních tvarů v korpusu SYN2000 (je jich 1 763 818 – viz výše) tedy můžeme říci, že plných 47 % všech tvarů se vyskytuje pouze jednou.

**5. Určete pořadí slovních druhů podle četnosti jejich výskytů:**

V následující tabulce jsou seřazeny slovní druhy podle četnosti jejich výskytů v korpusu SYN2000. Ve druhém sloupci je uveden celkový počet výskytů daného slovního druhu v korpusu, ve třetím je totéž vyjádřeno v procentech, dále je celkový počet výskytů daného slovního druhu v FSDT a konečně v posledním sloupci jeho vyjádření v procentech.

| Slovní druh | Frekvence SYN2000 | % SYN2000 | Frekvence FSDT | % FSDT |
|---|---|---|---|---|
| podstatná jména | 32 479 357 | 32,48 % | 450 798 | 27,77 % |
| slovesa | 14 930 501 | 14,93 % | 294 726 | 18,15 % |
| přídavná jména | 12 170 540 | 12,17 % | 181 229 | 11,16 % |
| předložky | 10 596 059 | 10,60 % | 164 395 | 10,12 % |
| zájmena | 8 739 530 | 8,74 % | 177 155 | 10,91 % |
| spojky | 6 967 738 | 6,97 % | 158 718 | 9,78 % |
| příslovce | 6 369 074 | 6,37 % | 167 008 | 10,29 % |
| číslovky | 4 945 700 | 4,95 % | 26 107 | 1,61 % |

| | | | | |
|---|---|---|---|---|
| částice | 550 029 | 0,55 % | - | - |
| citoslovce | 26 486 | 0,03 % | 3 391 | 0,21 % |

Pokud sečteme všechny frekvence výskytů v korpusu, dostaneme se k číslu 97 775 014 (tj. 97,78 %). Zbytek do 100 miliónů, přesněji řečeno 2 286 371 výskytů, tvoří slovní druhy, které z nejrůznějších důvodů (málo frekventovaná slova, vlastní jména, překlepy) nerozpoznal program pro automatické přiřazování lemmat a morfologických kategorií. K tomu je třeba ještě připočítat 20 847 328 výskytů různých interpunkčních znamének, takže celkový počet pozic v korpusu SYN2000 je 120 908 713. Pozicí zde rozumíme libovolný řetězec znaků, který lze samostatně vyhledávat, tedy nejen všechna slova, ale i čísla a interpunkční znaménka. Pokud jste někdy zkusili v našem korpusovém manažeru GCQP zadat v menu volbu Korpus > Souhrnné informace a zmátl vás udávaný počet pozic (120 908 713), vězte, že je všechno v nejlepším pořádku – skutečných slov je v korpusu SYN2000 skutečně „pouze" sto miliónů.

**6. Seřaďte interpunkční znaménka podle frekvence (od nejčastějšího k nejméně častému):**

| | |
|---|---|
| čárka | 7 647 971 |
| tečka | 7 201 853 |
| uvozovky | 1 530 104 |
| pomlčka | 1 220 801 |
| dvojtečka | 841 067 |
| uzavírací závorka | 800 972 |
| otevírací závorka | 769 185 |
| otazník | 260 904 |
| vykřičník | 134 301 |
| lomítko | 127 555 |

Jak již bylo řečeno výše, všech interpunkčních znamének je v korpusu SYN2000 celkem něco přes 20 miliónů, dá se tedy říci, že zhruba za každým pátým slovním tvarem následuje interpunkční znaménko.

**7. Určete pořadí nejčastějších předložek podle frekvence:**

| | |
|---|---|
| v/ve | 2 525 070 |
| na | 1 664 860 |
| s/se | 886 521 |
| z/ze | 861 027 |

| | |
|---|---|
| o | 660 813 |
| do | 621 814 |
| k/ke | 523 095 |
| pro | 401 812 |
| za | 391 733 |
| po | 319 788 |

**8. Určete pořadí nejvíce frekventovaných spojek od nejčastější k nejméně časté:**

| | |
|---|---|
| a | 2 831 331 |
| že | 867 283 |
| i | 543 845 |
| ale | 389 016 |
| aby | 205 966 |
| nebo | 201 500 |
| když | 193 081 |
| však | 190 568 |
| až | 162 687 |
| než | 153 767 |

**9. Které podstatné jméno se vyskytuje nejčastěji?**

| Pořadí SYN2000 | Frekvence SYN2000 | Pořadí FSDT | Pořadí FSP |
|---|---|---|---|
| rok | 406 834 | pán/pan | rok |
| člověk | 183 509 | práce | práce |
| Praha | 131 958 | ruka | lidé |
| strana | 130 835 | zem/země | léta |
| doba | 111 053 | lidé | strana |
| den | 110 289 | doba | děti |
| země | 103 622 | hlava | organizace |
| společnost | 90 657 | oko | úkol |
| svět | 85 503 | cesta | doba |
| firma | 84 285 | chvíle | den |

V tabulce vidíme v prvním a druhém sloupci nejčastější substantiva a jejich frekvence v SYN2000. V následujících sloupcích je prvních 10 nejčastějších substantiv ze dvou frekvenčních slovníků: FSDT a Těšitelová, M. a kol.:

Frekvenční slovník současné české publicistiky. Praha 1980 (FSP). Z tohoto srovnání je patrné, že vysoké zastoupení publicistiky v korpusu SYN2000 značně ovlivňuje frekvenční údaje a nejčastějšími slovy v celém korpusu se tak stávají ta, která jsou nadměrně užívána v publicistických textech (*rok, strana, doba*). Zároveň je možné sledovat i dobové proměny publicistiky, např. Praha je v FSP uvedena až jako 21. substantivum, kdežto v korpusu zaujímá 3. místo. Ještě nápadnější je to u adjektiv: v SYN2000 je nejfrekventovanější adjektivum *český* a ve FSP *socialistický*.

**10. Které mužské jméno je nejvíce frekventované?**

| Pořadí SYN2000 | Frekvence SYN2000 | Pořadí FSDT | Frekvence FSDT |
|---|---|---|---|
| Jan | 41 570 | Václav | 660 |
| Jiří | 34 506 | Jan | 382 |
| Václav | 28 125 | Lojzík | 369 |
| Petr | 26 018 | Ota/Oto | 335 |
| Josef | 25 226 | Frantík | 272 |
| Pavel | 21 384 | Boleslav | 235 |
| Karel | 19 641 | Jaroslav | 226 |
| Vladimír | 16 650 | Karel | 187 |
| Jaroslav | 16 062 | Jindřich | 170 |
| Martin | 14 736 | Josef | 168 |

Problémem u většiny těchto jmen je homonymie některých jejich tvarů s jejich ženskými protějšky, což někdy vede k nesprávné lemmatizaci. Navíc je to homonymie velice častá, neboť se může týkat téměř všech tvarů singuláru. Výše uvedená čísla jsou tedy z tohoto pohledu potenciálně nespolehlivá a je vhodné se ještě před jejich nekritickým přijetím pokusit o jejich upřesnění.

Zkoumejme nyní slovo *Jan*. Homonymními tvary, které mohou frekvenci lemmatu *Jan* ovlivňovat, jsou *Jan, Jana, Janu, Jani* a *Jany*. Specifickým případem je tvar Janů, který je homonymní s příjmením známé zpěvačky, což nemusí tagger (program pro automatické určování lemmat a přiřazování morfologických kategorií) vědět a tím vlastně všechny výskyty tvaru *Janů* označit bez nejmenších pochybností jako genitiv plurálu slova *Jan*, což by mohlo právě díky očekávané vysoké frekvenci výskytů tohoto příjmení významně ovlivnit statistiku. Podobně je třeba dávat pozor na tvar *Jane*, který je ve většině případů nominativem anglického ženského jména. Na těchto příkladech je vidět, že při prověřování je nutné brát v úvahu i tvary, které existují spíše teoreticky

a v jazyce se s nimi prakticky nesetkáme. Následující tabulka ukazuje v prvním číselném sloupci celkový počet výskytů každého tvaru (frekvence tvaru), druhý sloupec udává, kolika z nich bylo jako lemma přiřazeno slovo *Jan* (# lemma Jan), třetí sloupec vyjadřuje tento poměr v procentech (% lemma Jan) a konečně čtvrtý sloupec ukazuje, jaký by tento poměr měl být, pokud by bylo lemma přiřazeno správně (% skutečnost). Zatímco čísla v prvních třech číselných sloupcích jsou přesná (přesněji řečeno ve třetím jsou zaokrouhlená), k procentům ve čtvrtém sloupci dojdeme ruční kontrolou dostatečně velkého náhodného vzorku, který nám vybere korpusový manažer.

| | Tvar | Frekvence tvaru | # lemma Jan | % lemma Jan | % skutečnost |
|---|---|---|---|---|---|
| 1. | Jan | 25 951 | 25 833 | 100 % | 100 % |
| 2. | Jana | 14 345 | 9 273 | 65 % | 64 % |
| 3. | Janu | 1 249 | 833 | 67 % | 80 % |
| 4. | Jani | 11 | 11 | 100 % | 0 % |
| 5. | Jany | 732 | 2 | 0 % | 0 % |
| 6. | Janů | 216 | 216 | 100 % | 8 % |
| 7. | Jane | 997 | 997 | 100 % | 6 % |

Vidíme, že v řádcích 1, 2 a 5 odpovídá očekávaný počet výskytů výsledkům lemmatizace a ve 4. řádku je rozdíl v absolutních číslech zanedbatelný. Zbývají tedy řádky číslo 3, 6 a 7. Ve 3. řádku je rozdíl mezi očekávaným počtem výskytů a výsledkem lemmatizace 13 % z 1 249, tedy je třeba k celkovému počtu výskytů lemmatu Jan přičíst asi 160 výskytů, označených jako femininum. Naopak podle 6. řádku odečteme 92 % z 216, čili asi 200, a podle 7. řádku dokonce 94 % z 997, tedy asi 940. Vidíme tedy, že skutečný počet výskytů lemmatu Jan je zhruba o 1000 nižší, než kolik ukazuje současná lemmatizace. Rozdíl to ale není velký, jde přibližně o 2,5 %, a na celkovém pořadí se nic nezmění, „nedožene-li" ovšem lemma *Jan* nějaké jiné, s původně nižší frekvencí. Podle našich propočtů se však pořadí lemmat mužských jmen nezmění, a proto jej můžeme považovat za platné.

Z výše uvedeného je patrné, že při práci s korpusem je třeba vědět něco o způsobu, jak se korpus tvoří, s jakými problémy se jeho tvorba potýká a brát tyto věci v úvahu. Není vždy dobré pouze bezmyšlenkovitě přebírat čísla či kontexty bez jejich ověřování jinými metodami. Pro určení pořadí jednotlivých jmen podle frekvence není samozřejmě nutné složitě provádět takovéto výpočty, které navíc znamenají v podstatě ignorování lemmatizace a taggování (přiřazení

morfologických kategorií); pokud si nečiníme nároky na přesnost, je možné k podobnému závěru dojít daleko dříve. Na druhou stranu však není moudré ihned zavrhovat značkování korpusu jenom proto, že jsou v něm chyby. V případě slov s mnoha různými tvary s nízkou homonymií (například sloveso *být*) je značkování korpusu neocenitelné. O tomto poměrně obtížném případu navíc nelze říci, že v současnosti používaný statistický přístup k taggování a lemmatizaci zcela selhal, chyby nebyly nijak zvlášť významné. Co říci závěrem této úvahy? Snad jen to, že korpus je pouhým (více či méně dokonalým) nástrojem, který je nutné umět používat; myslet a vyvozovat závěry musí zase jenom člověk, to za něj stroj (doufejme, že ani v budoucnu) neudělá.

**11. Vyberte nejčastější ženské jméno:**

| Pořadí SYN2000 | Frekvence SYN2000 | Pořadí FSDT | Frekvence FSDT |
|---|---|---|---|
| Marie | 9 070 | Blažena | 492 |
| Jana | 7 679 | Stázka | 465 |
| Eva | 5 839 | Marie | 250 |
| Anna | 4 260 | Jarmila | 196 |
| Hana | 3 934 | Olga | 190 |
| Helena | 3 343 | Marta | 177 |
| Věra | 3 325 | Markétka | 174 |
| Petra | 3 307 | Kateřina | 126 |
| Kateřina | 3 218 | Jarka | 113 |
| Zuzana | 3 121 | Tonička | 84 |

Homonymie některých tvarů jmen *Jana* a *Petra* s jejich mužskými protějšky *Jan* a *Petr* byla podrobně rozebrána u předchozího bodu. Také skutečné pořadí ženských jmen zhruba odpovídá pořadí v tabulce (určenému lemmatizací), na prvních místech se nic nezmění, pouze jméno *Petra* se vytratí z první desítky, protože jeho skutečná frekvence je nižší, než udává tabulka. Pro srovnání uvádíme ve vedlejší tabulce nejčastější ženská jména podle FSDT. Na první pohled je patrný vliv výběru beletrie (vysoká frekvence jmen *Blažena* a *Stázka*).

**12. O které zemi se v korpusu nejvíce píše?**

| | |
|---|---|
| ČR | 46 607 |
| USA | 45 576 |
| Německo | 28 744 |
| Rusko | 20 694 |

| | |
|---|---|
| Slovensko | 20 353 |
| Francie | 16 591 |
| Polsko | 13 037 |
| Itálie | 12 961 |
| Rakousko | 10 725 |
| Československo | 9 788 |

Jak je vidět, má naše země v této tabulce velkou konkurenci: jsou jí Spojené státy. Právě tvar *Spojené státy* má frekvenci rovných 9 100, což je téměř stejně jako *Československo*, a tak jenom potvrzuje, že USA patří jasné druhé místo těsně za ČR.

Pro zajímavost: tvar *Česko* se vyskytuje 2 854 krát, z toho více než třetina je součást složenin typu *Česko-německý*. *Česko* se začalo v publicistice objevovat ve větší míře již v roce 1993 v souvislosti s diskusemi o názvu nového státu, ale najdeme ho i v daleko starší beletrii, např. ve Škvoreckého Příběhu inženýra lidských duší.

**13. O kterém kontinentu se naopak píše nejméně?**

| | |
|---|---|
| Antarktida | 428 |
| Austrálie | 3 520 |
| Asie | 3 787 |
| Afrika | 4 199 |
| Amerika | 10 030 |
| Evropa | 41 121 |

**14. Které zahraniční město je jmenováno nejčastěji?**

| Pořadí SYN2000 | Frekvence SYN2000 | Frekvence FSDT |
|---|---|---|
| Moskva | 12 984 | 58 |
| Paříž | 10 022 | 113 |
| Bratislava | 9 224 | 38 |
| New York | 9 153 | 11 |
| Londýn | 8 321 | 27 |
| Washington | 6 994 | 10 |
| Vídeň | 6 522 | 97 |
| Berlín | 5 157 | 35 |
| Brusel | 5 019 | 7 |
| Řím | 4 753 | 54 |

Ve všech případech jde poměrně často o publicistické užití typu „Bratislava (ČTK) – Mluvčí slovenského premiéra..." nebo používání názvu hlavního města pro označení vlády celých zemí („...dává Moskvě signál..."). Právě toto užití se v publicistice pro *Moskvu* nejvíce vžilo, což by mohlo vysvětlovat její nejvyšší frekvenci. Zajímavé je, že dvanácté nejfrekventovanější slovo, které nacházíme v okolním kontextu slova *Moskva*, je Procházková.

Naše pořadí se značně liší od pořadí v FSDT uvedeného v druhém sloupci. Domníváme se, že je to způsobeno velkým zastoupením novin v korpusu a také změnou statutu některých měst (např. *Bratislava* je pro nás hlavním městem sousedního státu, a tak je uváděna ve zpravodajstvích – viz výše; *Brusel* se stal hlavním městem EU a píše se o něm častěji). Naproti tomu *Paříž* a *Vídeň* se daleko více vyskytují v beletrii a u Paříže je nápadný i její větší výskyt v časopisech, které píší o módě. Poměrně často je zmiňován i závod *Paříž–Dakar*, ovšem na první místo to nestačí. Další zvýšení frekvence u všech hlavních měst zemí plyne ze zpráv o počasí. Původní nepočeštěné názvy jako např. *London, Paris* nebo *Wien* mají jen malou frekvenci (okolo 5 % výskytů), navíc jde často o užití v cizojazyčných citacích nebo adjektivní užití v cizojazyčných názvech („London Business School", „The Paris Review"), a proto jsme tyto tvary nevzali v úvahu.

**15. O kterém zahraničním městě, které však není hlavním městem, se nejvíce píše?**

| | |
|---|---|
| New York | 9 153 |
| Sarajevo | 3 955 |
| Mnichov | 3 687 |
| Jeruzalém | 3 431 |
| Barcelona | 2 872 |
| Los Angeles | 2 759 |
| Košice | 2 369 |
| Chicago | 2 283 |
| Frankfurt | 2 262 |
| Atlanta | 2 109 |

Na tomto seznamu je dobře patrné, že odhadnout dopředu alespoň přibližně pořadí zmíněných měst je téměř nemožné, protože frekvence každého z nich ovlivňují jiné faktory. První místo *New Yorku* se dalo očekávat, kdo by si však vzpomněl na *Sarajevo*? Jeho druhá pozice je samozřejmě odrazem zpravodajství o válce v bývalé Jugoslávii, nikoli atentátem v roce 1914 ani „pučem"

o třiaosmdesát let později. Podobně nepřekvapí ani čtvrté místo *Jeruzaléma*. Třetímu místu *Mnichova* samozřejmě pomohly události z roku 1938, nejvíce výskytů je ale přesto z oblasti sportu – Bayern Mnichov, olympiáda v roce 1972. V této souvislosti překvapí jedna z nejčastějších kolokací slova Mnichov – a to číslovka 1860. Že si nepamatujete, co se tehdy stalo, natož v Mnichově? Nestalo se možná nic významného, pouze se tak jmenuje mnichovský fotbalový klub „Mnichov 1860". Podobně je na tom i *Barcelona* – letní olympijské hry v roce 1992 a samozřejmě FC Barcelona „zařídily" tomuto městu páté místo a frekvenci vyšší, než má například Madrid. Kam se hrabe Real na „Barcu"!

Frekvence *Los Angeles* je relativně málo ovlivněna sportem, jde přeci o významné americké město, přesto se v konkordancích objeví dost zmínek o hokeji, olympiádě v roce 1984, seriálu „Policejní okrsek Los Angeles" apod. Struktura rozložení výskytů města *Košic* je podobná spíše městům českým než zahraničním, jedním z důvodů je také většina výskytů ve sportovním zpravodajství. *Chicago* je na tom podobně jako Los Angeles, tedy poměrně „zasloužená" pozice, relativně málo ovlivněná sportem. *Frankfurt* je na tom ještě lépe: sportu je tady jen okolo 20 %, což je skutečně pěkný výsledek. *Atlanta* pak jako dějiště nedávných olympijských her samozřejmě „nezklame" četností výskytů z oblasti sportu.

**16. Která země se nazývá „země orlů"?**

Správná odpověď je Albánie. Pokud nevíte, proč se jí říká právě takto, vězte, že podle Respektu č. 43 z roku 1991 „Shqiperia (Albánie) znamená v albánštině „země orlů"", ačkoli Martin Novák v Nedělní příloze Lidových novin č. 14 roku 1993 píše, že tam „neviděl ani jednoho". Na tomto příkladu je vidět, že korpus lze kromě vyhledávání jazykových informací používat také jako encyklopedii, byť si v tomto ohledu nečiní pražádný nárok na úplnost či reprezentativnost a pravdivost si samozřejmě musíte také pohlídat sami.

**17. Které naše město se po Praze a Brně nejčastěji uvádí?**

| | |
|---|---|
| Praha | 131 958 |
| Brno | 36 976 |
| Plzeň | 20 712 |
| Ostrava | 19 129 |
| Prostějov | 16 438 |
| Olomouc | 16 291 |
| Hradec Králové | 11 015 |
| Zlín | 10 920 |

| | |
|---|---|
| Liberec | 9 212 |
| Pardubice | 9 079 |

Na této tabulce není nic překvapivého, snad jen vyšší frekvence výskytů názvu města Prostějov, kterou lze ovšem vysvětlit tím, že byly do korpusu SYN2000 zařazeny regionální noviny Prostějovský týden.

**18. Který ze dnů v týdnu je nejfrekventovanější?**

| Pořadí SYN2000 | Frekvence SYN2000 | Upravená frekvence | Frekvence FSDT |
|---|---|---|---|
| sobota | 21 919 | 21 592 | 82 |
| neděle | 19 859 | 19 855 | 251 |
| pátek | 13 485 | 13 580 | 64 |
| pondělí | 12 817 | 12 800 | 52 |
| středa | 12 551 | 12 505 | 55 |
| čtvrtek | 10 113 | 10 429 | 37 |
| úterý | 10 341 | 10 329 | 36 |

U dnů v týdnu nastává podobná potíž jako u výše uvedených vlastních jmen – homonymie. Názvy dnů v týdnu jsou homonymní s příjmeními (např. *Čtvrtek, Pátek, Sobota*), názvy obcí (např. *Rimavská Sobota, Úterý*) a také se vyskytují v názvech literárních a filmových děl. Všechny tyto možnosti je třeba vzít v úvahu při zjišťování frekvence. Podívejme se na jednotlivé dny podrobněji.

Nejfrekventovanějším dnem v korpusu SYN2000 je ***sobota***. V konkordancích tohoto lemmatu také najdeme nejvíce případů, kdy toto slovo označuje něco jiného než den v týdnu. Jsou to: *Rimavská Sobota* (169 výskytů), *Luděk Sobota* (89 výskytů), ostatní výskyty jako příjmení 69. Celkově je tedy třeba odečíst 327 výskytů, ale jak vidíme ve výše uvedené tabulce, nestačí to k tomu, aby *sobota* přišla o své prvenství.

Druhým nejfrekventovanějším dnem v týdnu je ***neděle***. Ta se překvapivě prakticky nevyskytuje jinak než jako název dne, takže celkově přijde pouze o 4 výskyty: dvakrát je zmíněn film *Neděle na venkově* a dvakrát rozhlasová hra Daniely Fischerové *Neděle*.

Třetím dnem v našem frekvenčním pořadí je ***pátek***. V tomto případě nebudeme jen odčítat, ale i přičítat, protože pro SYN2000 existuje samotné lemma *Pátek*,

které zahrnuje především příjmení, jména Robinsonova druha a jeho jmenovců, ale část těchto výskytů (183) znamená den v týdnu. Jiné významy než den v týdnu: *magazín LN Pátek* (52 výskytů), obec *Pátek nad Ohří* (5 výskytů), Čapkovy *Pátky* – setkávání s prezidentem Masarykem a dalšími osobnostmi (7 výskytů), příjmení (2 výskyty), film *Pátek třináctého* (slovenská komedie – 1 výskyt, horor USA – 8 výskytů), rozhlasový pořad *Pátky s mikrofonem* (2 výskyty), jméno *Pátka* v Nohavicově písni *Robinzon* (7 výskytů), divadelní představení *Čulibrkárna aneb Pátek třináctého* (1 výskyt), obraz *Dopoledne Velkého Pátku* (1 výskyt), kniha *Pátek aneb lůno Pacifiku* (1 výskyt). Celkově tedy odečteme 95 výskytů, což nijak neovlivní třetí místo.

Dalším dnem v týdnu podle frekvence je ***pondělí***. Tady dojde také k zanedbatelnému odpočtu 17 výskytů: devětkrát je v korpusu zmíněna výtvarná skupina *Pondělí*, dvakrát obraz *Bohyně Pondělí*, dvakrát Hrabalovo *Modrý pondělí* a čtyřikrát je zmínka o zvyklosti *modré pondělí* (zvyk některých řemeslných cechů nepracovat v pondělí, původně snad z neschopnosti po „nedělní oslavě").

Následuje ***středa***. Všechny výskyty kromě názvů dnů v týdnu jsou příjmení – 46 výskytů. Najdeme mezi nimi například sochaře *Jiřího Středu* a brněnského jezuitu *Martina Středu*, proslulého svou statečností při obléhání Brna za třicetileté války.

Dalším pořadím ovšem podrobnější prozkoumání konkordancí přece jen zamíchalo. Z posledního místa na předposlední se dostává ***čtvrtek***, který od lemmatu „Čtvrtek" získal 332 výskytů, kdy nešlo o příjmení (např. *Václava Čtvrtka*), ale skutečně o den v týdnu. Odečteme však 16 výskytů: dvakrát příjmení, třináctkrát genitiv plurálu od lemmatu „čtvrtka" (velmi nepříjemná homonymie!) a jednou obec Spišský Čtvrtek. Celkový zisk 316 výskytů stačí na získání předposledního místa.

Posledním, nikoliv však málo zajímavým dnem je ***úterý***. V jeho konkordancích lze dvanáctkrát objevit západočeské městečko *Úterý*, kde se koná veřejný závod na horských kolech *Úterské dršky*.

Nakonec ještě malý přehled přívlastků, které jsou používány jako názvy pro konkrétní dny v roce: *Velký pátek* (119 výskytů), *Bílá sobota* (25 výskytů), *Popeleční středa* (37 výskytů), *Květná neděle* (31 výskytů), *Zelený čtvrtek* (20 výskytů).

A na úplný závěr několik českých a slovenských obcí „na celý týden": Pondelok, Úterý, Streda nad Bodrogom, Spišský Štvrtok, Pátek nad Ohří,

Sobotka, Neděliště. To jsou místa, která byla zvolena pro výzkum, jak se za určité období změnil život v ČR a na Slovensku – kritériem byl název podle dnů v týdnu (Reflex, číslo 34 z roku 1994). To všechno se můžete dočíst v korpusu.

**19. Který měsíc má nejvíce výskytů?**

| SYN2000 | | CNK2 | | FSDT | |
|---|---|---|---|---|---|
| září | 18 891 | leden | 23 811 | září | 204 |
| leden | 18 541 | září | 22 137 | říjen | 78 |
| květen | 18 197 | listopad | 20 152 | duben | 69 |
| listopad | 17 812 | říjen | 19 311 | květen | 67 |
| červen | 17 341 | květen | 19 022 | únor | 66 |
| říjen | 16 604 | červen | 18 679 | srpen | 66 |
| duben | 15 380 | prosinec | 17 823 | listopad | 60 |
| březen | 15 190 | březen | 16 996 | březen | 58 |
| srpen | 14 806 | duben | 16 801 | leden | 50 |
| prosinec | 14 383 | srpen | 16 259 | červen | 39 |
| červenec | 14 048 | červenec | 16 101 | červenec | 37 |
| únor | 13 663 | únor | 15 752 | prosinec | 31 |

V tabulce vidíme pořadí měsíců podle frekvence. V prvním sloupci je pořadí a frekvence z korpusu SYN2000, ve druhém z bývalého korpusu CNK2, který obsahoval 120 mil. slovních tvarů a 80 % textů tvořila publicistika, ve třetím sloupci jsou pak frekvence z FSDT. Rozdíly mezi nimi jsou však nepatrné, protože jde o nízký počet výskytů. Vidíme, že pořadí měsíců v obou korpusech (SYN2000 a CNK2) se výrazněji neliší, což přičítáme relativně vysokému zastoupení publicistiky. Domníváme se, že vysoký počet výskytů *září* v FSDT je způsoben tím, že většina jeho publicistických textů pochází právě z tohoto měsíce.

**20. Které číslo je nejčastější (psané číslicí)?**

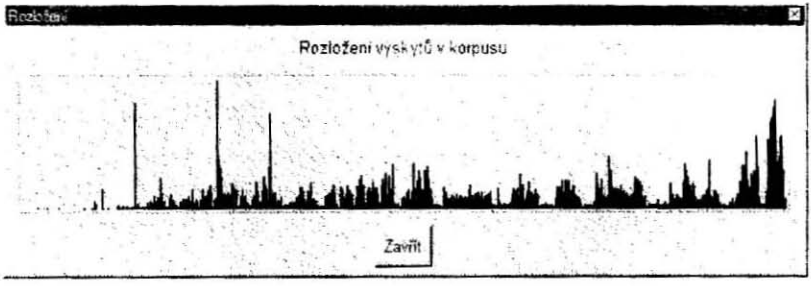
| | |
|---|---|
| 1 | 219 999 |
| 2 | 175 084 |
| 3 | 141 289 |
| 6 | 107 728 |

| | |
|---|---|
| 4 | 103 826 |
| 5 | 85 077 |
| 0 | 84 577 |
| 7 | 66 226 |
| 10 | 63 337 |
| 8 | 55 064 |
| 9 | 45 670 |

Až neuvěřitelná je shoda ve tvarech následujících grafů rozložení u číslovek psaných číslicí (vodorovná osa je rozdělena na malé úseky, z nichž každý odpovídá určitému úseku textu v korpusu; počet výskytů daného čísla v každém z nich je pak vynesen vertikálně). Tato shoda se samozřejmě dá vysvětlovat tím, že čísla se nejčastěji vyskytují v nejrůznějších tabulkách (pořadí na výsledkových listinách, skóre ve sportovních výsledcích, čísla tahů v šachových partiích, pořadová čísla stranických sjezdů, míry, váhy, ceny atd.), které se pochopitelně vyskytují stále na stejném místě v korpusu. Gólů však nepadá neomezený počet, každá výsledková listina či šachová partie někdy končí a žádná strana netrvá věčně. Právě proto je pozoruhodné, že se i při velkém úbytku počtu výskytů stále udržuje až na detaily tentýž tvar grafu. Můžete namítnout, že tato podobnost jistě někde končí. Ano, ale vytrácí se jenom velmi zvolna, což je navíc způsobeno také odlišným zobrazením grafu – je-li totiž výskytů jen pár tisíc, vzrůstá jeho nespojitost. Na důkaz uveďme náhodně vybranou číslovku 37, jež má pouhých 7549 výskytů. Přesto je však její graf podobný číslovce 1, která je ovšem v korpusu takřka třicetkrát frekventovanější!

Obdobně velkou podobnost lze vysledovat také v otázce č. 22 (denní hodina), zatímco rozložení jednotlivých roků v následující otázce č. 21 již tak významnou podobnost nevykazuje.

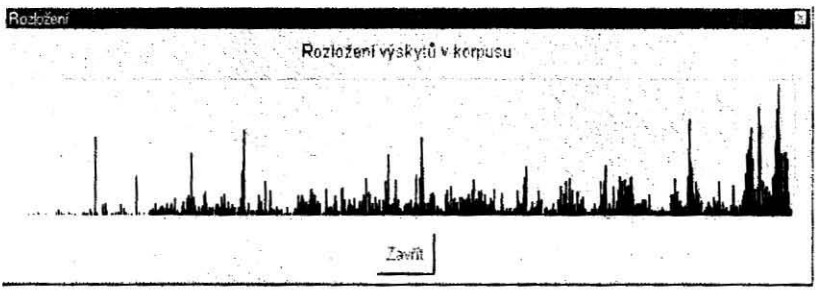
*1*



*2*

*3*

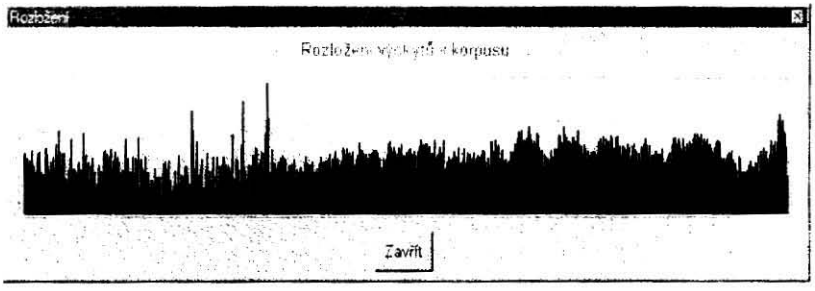
*37*




Pro zajímavost uvádíme ještě frekvence číslovek psaných slovy. Není jistě žádným překvapením, že jejich rozložení (na rozdíl od číslovek psaných čísly), je úplně jiné, a to daleko rovnoměrnější. I když jsou si grafy číslovek psaných slovy také poměrně podobné, není tato podobnost taková jako v případě číslovek psaných čísly. Následující graf ukazuje rozložení výskytů číslovky *dva*.

| | |
|---|---|
| jeden | 211 586 |
| dva | 153 114 |
| tři | 86 111 |
| čtyři | 43 971 |
| pět | 42 267 |
| deset | 31 196 |
| šest | 25 201 |
| sedm | 18 713 |
| osm | 16 995 |
| devět | 10 614 |
| nula | 2 852 |

***dva***

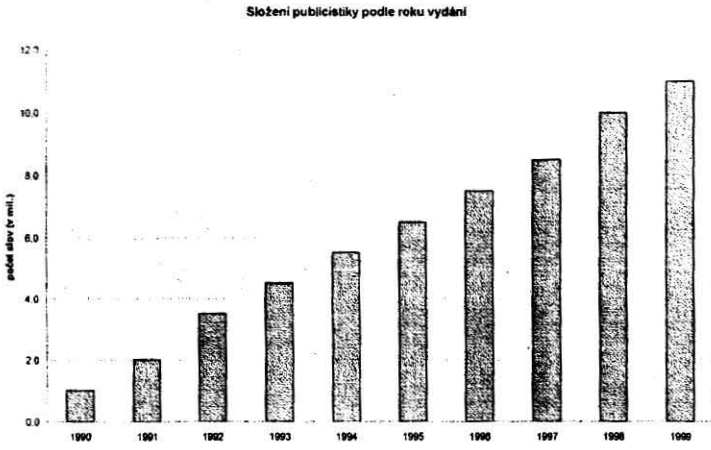



**21. Který rok je nejhojněji zastoupen? (připomínáme, že v korpusu převažují texty z let 1990–1999)**

| | | | | | |
|---|---|---|---|---|---|
| 1. | 1994 | 16 862 | 12. | 1999 | 8 355 |
| 2. | 1995 | 16 255 | 13. | 1968 | 4 625 |
| 3. | 1996 | 15 911 | 14. | 1988 | 3 885 |
| 4. | 1997 | 15 576 | 15. | 1948 | 3 462 |
| 5. | 1993 | 14 391 | 16. | 1945 | 3 389 |
| 6. | 1992 | 13 184 | 17. | 1987 | 2 641 |
| 7. | 1998 | 12 908 | 18. | 1986 | 2 541 |
| 8. | 1991 | 10 486 | 19. | 1984 | 2 532 |
| 9. | 1990 | 10 431 | 20. | 1980 | 2 507 |
| 10. | 1989 | 10 134 | 21. | 1985 | 2 364 |
| 11. | 2000 | 9 486 | 22. | 1969 | 2 167 |

Podíváme-li se na tabulku, vidíme, že rok 1994 je jakýmsi nejfrekventovanějším středem, za kterým následuje vzestupná řada let 1995, 1996 a 1997, dále jsou roky 1993 a 1992, což je naopak sestupná řada začínající rokem 1994, následuje rok 1998 (pokračování vzestupné řady) a léta 1991, 1990 a 1989 (opět sestupná řada). Rok 2000 předstihl rok 1999 nejenom proto, že se o něm často mluví jako o přelomu milénia („...v nynější formě rok 2000 přežijí..."), ale hlavně z toho důvodu, že jde o číslovku často používanou v nejrůznějších odhadech („...stojí ho to maximálně 2000 Kč."). Pokračování sestupné řady 1988, 1987 atd. je přirozené, prokládané je pouze významnými roky naší historie (1968, 1948, 1945) a lety 1984 a 1980. Proč jsou právě tyto dva roky významné? Nyní by nás již nemělo překvapit, že ještě významnějším faktorem než Orwellův román je skutečnost, že se tehdy konaly olympijské hry.

To je všechno logické a pochopitelné. Překvapující ale je, že se v korpusu nejvíce vyskytuje právě rok 1994. Složení publicistiky, která toto pořadí nejvíce ovlivňuje, je totiž takové, aby co nejvíce odráželo současný jazyk; nejvíce jsou tedy zastoupeny texty z roku 1999 a nejméně z roku 1990, jak je vidět na následujícím grafu. Vezmeme-li v úvahu, že lépe jsou na tom spíše roky ze středu devadesátých let (je možné na ně odkazovat jak ze starších, tak i z novějších novin) než roky okrajové (např. 1999), a dále fakt, že publicistika spíše odkazuje nazpět než dopředu (píše více o tom, co se již stalo), vychází nám jako ideální kandidát některý z roků 1996 nebo 1997. Prvenství roku 1994 tak zůstává malým korpusovým tajemstvím.



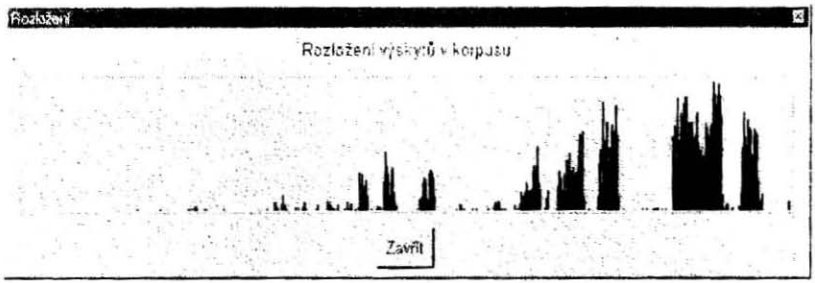
**22. Která denní hodina je v korpusu nejčastější?**

| | |
|---|---|
| 20.00 | 16 017 |
| 19.30 | 10 788 |
| 17.30 | 9 836 |
| 19.00 | 9 183 |
| 17.00 | 8 059 |
| 18.00 | 7 148 |
| 15.00 | 5 375 |
| 16.30 | 5 145 |
| 9.00 | 4 929 |
| 14.00 | 4 633 |

Zamyslíme-li se nad tím, odkud se tato čísla v korpusu berou a co tedy určuje jejich pořadí, asi nás hned napadne, že jde hlavně o programy – ať už televizní a rozhlasové nebo programy kin, divadel, klubů a podobně. Právě ve 20.00 začíná hlavní televizní program a většina představení. A pokud ne ve 20.00, tedy v 19.30. Třetí místo času 17.30 vysvětluje fakt, že většina kin začíná hrát odpoledne právě od 17.30. Překvapením tedy může být snad jen umístění 9.00 v první desítce – opět se podíváme do korpusu a vidíme, že to je hodina, kdy otevírá většina galerií a výstav...

Netypický, ilustrativní, a svým způsobem i pěkný je následující graf, ukazující rozložení časového údaje 20.00 v korpusu. Před tím, než ho začneme komentovat, bychom však rádi přesněji popsali, co je na něm vlastně vidět. Jak již bylo řečeno, vodorovná osa x je rozdělena na mnoho malých úseků, z nichž každý odpovídá určitému úseku textu v korpusu. Počet výskytů daného slova v každém z nich je vynesen na osu y (vertikálně). Řazení těchto úseků pak odpovídá pořadí textů v korpusu, které je následující: nejvíce vlevo je beletrie (15 % celku), následuje odborná literatura (25 %) a publicistika (60 %). Vidíme, že časový údaj 20.00 se vyskytuje de facto výhradně v publicistice, v rámci které navíc tvoří příkré skalní útesy a kaňony, dané zřejmě umístěním příloh s programem televize. Jak již bylo naznačeno, velmi podobná rozložení mají také ostatní časové údaje.

*20.00*




**23. Jak jdou za sebou podle frekvence čtyři nejčastěji zmiňované politické strany?**

| | |
|---|---|
| ODS | 26 936 |
| ČSSD | 15 523 |
| ODA | 9 261 |
| KDU | 6 782 |
| ČSL | 6 258 |
| US | 4 039 |
| KSČM | 3 950 |
| KSČ | 3 184 |

Opět žádné velké překvapení, pouze je třeba pro upřesnění dodat, že téměř třetina výskytů zkratky US znamená „americký", dalším významem jsou potom „Uhelné sklady". Pokud bychom tyto výskyty odečetli, propadla by se US až na konec za KSČ, což je pochopitelné, neboť jde o stranu novou, zatímco KSČ má u nás poměrně dlouhou tradici.

**24. O kterých nositelích akademického titulu se nejvíce píše?**

| | | | | | |
|---|---|---|---|---|---|
| 1. | Ing. | 11 677 | 10. | RNDr. | 790 |
| 2. | Dr. | 9 007 | 11. | DrSc. | 349 |
| 3. | MUDr. | 2 764 | 12. | MVDr. | 221 |
| 4. | prof. | 2 554 | 13. | ThDr. | 102 |
| 5. | JUDr. | 2 318 | 14. | PaedDr. | 56 |
| 6. | CSc. | 1 633 | 15. | RSDr. | 31 |
| 7. | doc. | 1 351 | 16. | Bc. | 17 |
| 8. | PhDr. | 1 196 | 17. | PharmDr. | 15 |
| 9. | Mgr. | 1 026 | 18. | RCDr. | 5 |

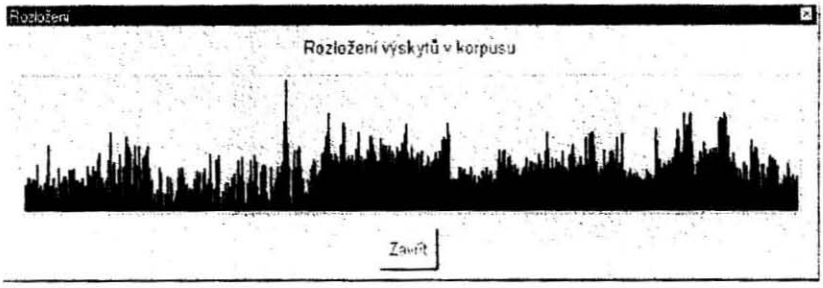
Asi vám nebude známý poslední titul RCDr. – doctor rerum commercialium čili doktor obchodních věd. Všech pět výskytů pochází ze studie Prokopa Závodského „Přehled vývoje vysokého ekonomického školství na našem území", který vyšel v Acta Oeconomica Pragensia 5/99. Tady se také dovídáme, že v roce 1919 byla založena Vysoká škola obchodní (VŠO) v Praze, i když o její zřízení současně usilovala Olomouc. Patřila pod ČVUT a v roce 1929 bylo schváleno, že může udělovat titul RCDr. První dva lidé obdrželi tento titul v roce 1935 a jejich promoce se zúčastnil dokonce sám prezident Masaryk. Do roku 1939 jej získalo 30 mužů, někteří z nich jsou dále v článku jmenováni.

**25. Jak jdou za sebou z hlediska frekvence:**

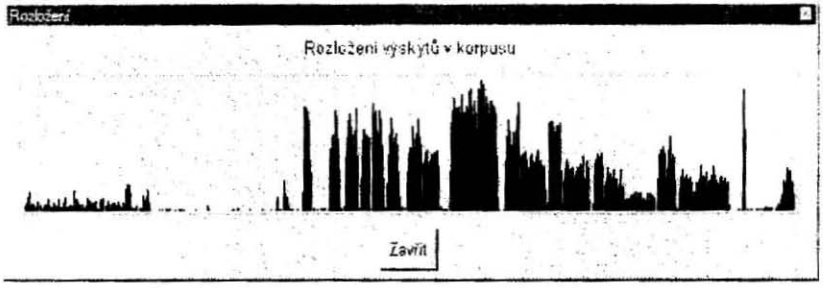
| | |
|---|---|
| dnes | 79 927 (včetně 4 399 výskytů tvaru *dneska*) |
| včera | 60 799 |
| zítra | 8 922 (včetně 283 výskytů tvaru *zejtra*) |

Na následujících grafech rozložení výskytů v korpusu je názorně vidět odlišný způsob používání těchto tří slov. Hranice jednotlivých typů textu (beletrie, odborné texty a publicistika), které tvoří jakési tři segmenty, jsou na grafech dobře zřetelné. Vidíme, že nejrovnoměrněji je rozloženo používání slova *dnes*, které se běžně používá i v odborných textech, většinou ve významu „v současné době". Hranice jednotlivých textových typů nejsou vidět. Naproti tomu se slova *včera* a *zítra* v tomto smyslu používají jen zřídka, a není tedy mnoho důvodů pro to, aby se v odborných textech objevovala častěji. Obě tato slova se používají v beletrii i publicistice, takže je předěl mezi těmito dvěma typy textu a odbornou literaturou dobře patrný. Zatímco je ale výskyt slova *zítra* mezi těmito dvěma segmenty (publicistikou a beletrií) rozdělen víceméně rovnoměrně, u výskytů slova *včera* je vidět velká převaha publicistiky – noviny přece píší spíše o tom, co se stalo, než o tom, co teprve bude. Nejpravděpodobnějším vysvětlením „děr" ve výskytu slova *včera* v publicistice jsou regionální noviny a jiné nedeníky, pro které používání slova *včera* ztrácí smysl. Tyto „díry" nejsou u slova **zítra** patrné proto, že nejde o (relativně) takový pokles jako u slova *včera*. Na závěr ještě dovolte poznámku pro uživatele našeho korpusu (kterým se můžete stát i vy!): tyto grafy rozložení jednotlivých slov získáte v našem korpusovém manažeru GCQP volbou menu Konkordance>Statistiky>Rozložení.

*dnes*




*včera*




*zítra*

**26. Seřaďte rozměry podle frekvence:**

| | |
|---|---|
| délka | 8 745 |
| výška | 7 968 |
| šířka | 4 425 |
| hloubka | 3 290 |

**27. Seřaďte následující adjektiva od nejfrekventovanějšího k nejméně frekventovanému (pod jedním lemmatem (=základní slovníkový tvar) jsou započteny i všechny jeho stupně):**

| | |
|---|---|
| vysoký | 78 148 |
| dlouhý | 35 717 |
| široký | 15 981 |
| hluboký | 10 941 |

Na první pohled je zajímavé, že ačkoliv *výška* není nejčastějším rozměrem, adjektivum *vysoký* obsadilo s velkým náskokem první místo; přitom pořadí i frekvence ostatních adjektiv přibližně odpovídají předchozí tabulce rozměrů. Vysvětlení nám napoví následující tabulka, ze které je patrný velký nepoměr v používání druhého a třetího stupně slova *vysoký* a ostatních adjektiv. Zatímco o všech ostatních lze říci, že poměr mezi frekvencemi 1., 2. a 3. stupně je přibližně stejný, v žádném případě to neplatí pro adjektivum *vysoký*. Právě nárůst v absolutních číslech, způsobený tímto nepoměrem, vysvětluje onen zmíněný velký náskok slova *vysoký*.

| | 1. stupeň | 2. stupeň | 3. stupeň |
|---|---|---|---|
| vysoký | 32 819 | 24 981 | 20 348 |
| dlouhý | 26 623 | 8 008 | 1 086 |
| široký | 10 378 | 4 632 | 971 |
| hluboký | 7 870 | 2 270 | 801 |

**28. Které z následujících slov se v korpusu nejčastěji vyskytuje s pravopisnou chybou?**

| | | |
|---|---|---|
| tamější | 589 správně, | 13 chyb |
| vzpomněl | 1976 správně, | 7 chyb |
| zapomněl | 2152 správně, | 7 chyb |
| domněnka | 239 správně, | 0 chyb |

K výše uvedeným číslům bychom rádi poznamenali, že chybný tvar *zapoměl* se v korpusu objevil celkem 14krát, v polovině případů však nešlo o chybu, ale o vlastní jméno nebo součást rozboru pravopisných chyb.

**29. Určete počet výskytů chybného tvaru *dvěmi*. Počet výskytů správného tvaru *dvěma* je 15 882.**

Počet výskytů chybného tvaru *dvěmi* je 116, poměr *dvěmi*:*dvěma* je tedy

přibližně 1:137. Podíváme-li se na rozložení tvaru *dvěma*, zjistíme, že je takřka rovnoměrné. Naproti tomu tvar *dvěmi* se poměrně často vyskytuje v odborné literatuře – že by měli právě vědci sklon k hyperkorektnosti? Podobně tomu je i u slova *nejoptimálnější*, které se v beletrii nevyskytuje vůbec. Naopak slovo *nejideálnější* proniklo do všech tří textových kategorií. Nejvíce těchto tvarů najdeme v publicistice – to ukazuje tabulka níže, ve které je procentuální zastoupení všech tří zmíněných slov v jednotlivých textových kategoriích. Pro úplnost dodejme počty výskytů ostatních „stupňů": *ideální* 4 707 výskytů, *ideálnější* 13 výskytů, *optimální* 2 923 výskytů, *optimálnější* 6 výskytů a další 3 nalezneme v Jazykovém koutku časopisu Vesmír z roku 1995, který se věnuje používání tohoto adjektiva. Stejně jako u číslovky *dvěma* i zde jsou správné tvary častější.

| | dvěmi | nejoptimálnější | nejideálnější |
|---|---|---|---|
| beletrie | 10 % | 0 % | 17 % |
| odborné texty | 43 % | 49 % | 17 % |
| publicistika | 47 % | 51 % | 66 % |

**30. Který z názvů následujících jazyků je v korpusu nejčastěji zmiňován?**

| | | | | | |
|---|---|---|---|---|---|
| 1. | čeština | 4 120 | 8. | španělština | 276 |
| 2. | angličtina | 3 478 | 9. | italština | 248 |
| 3. | němčina | 1 628 | 10. | řečtina | 225 |
| 4. | francouzština | 1 041 | 11. | polština | 196 |
| 5. | latina | 692 | 12. | arabština | 187 |
| 6. | ruština | 657 | 13. | hebrejština | 170 |
| 7. | slovenština | 407 | 14. | čínština | 154 |

*Marie Kopřivová a Michal Křen*

# Za jakých podmínek můžete pracovat s korpusem SYN2000?

Reprezentativní korpus současné psané češtiny SYN2000 je přístupný po internetové síti. Stejným způsobem (přes síť) se tedy můžete dozvědět vše potřebné také o podmínkách a technických možnostech práce s ním. Naše internetová adrese je:

**ucnk.ff.cuni.cz**

Samozřejmě nás můžete oslovit i přímo: zajít na Filozofickou fakultu do sklepních prostor, kde se nacházejí místnosti číslo 51, 52, 53 a uvnitř celý náš kolektiv. Na internetové stránce i v naší pracovně Vám nabídneme k vyplnění a k podpisu formulář smlouvy, na jehož základě Vám bude (zdarma) přiděleno přístupové heslo, platné po dobu jednoho roku (prodlužujeme na požádání). Z podpisu smlouvy o užití SYN2000 pro Vás vyplývají jen tři povinnosti: ochránit data před komerčním využitím, dodat do Ústavu Českého národního korpusu kopii práce, která s pomocí vyhledávání v korpusu SYN2000 vznikne, a v této práci korpus citovat.

V případě zájmu o komerční využití korpusu a výsledků práce na něm (organizací i jednotlivcem!) je nutné vstoupit do jednání s Filozofickou fakultou a uzavřít s ní smlouvu o placeném přístupu (který je také termínovaný).

Na naší internetové stránce najdete i rady technické, jak nainstalovat Váš počítač, aby Vám vyhledávání v korpuse fungovalo, aby se Vám dobře zobrazovala česká diakritika, a na koho se obrátit o pomoc. Najdete tam také manuál (návod k obsluze), který Vás naučí, jak používat náš „manažer" (software, s jehož pomocí je možné v korpusu vyhledávat).

Pokud se pohybujete na Filozofické fakultě, máte možnost si zjistit konzultační hodiny pracovníků ÚČNK, v začátcích Vaší práce s korpusem Vám mohou pomoci.

V mezipatře, vlevo od průchodu do velké auly, je prostor, ve kterém jsme zřídili svou vlastní korpusovou pracovnu. Pomocí počítačů zde můžete pracovat se SYN2000 od sedmi hodin ráno do sedmi hodin večer. Ke vstupu do ní potřebujete jen kartu, kterou Vám za vratnou zálohu vydá v místnosti č. 51 dr. Blatná.

*Věra Schmiedtová*

# Obsah

Na úvod poděkování a vzpomínka (Věra Schmiedtová) ... 4

## POHLED DO KORPUSU

Island (Renata Blatná) ... 5
Zdrobněliny v korpusu (Jana Klímová) ... 6
Jazyková animace (Renata Blatná) ... 8
Obuv, boty, střevíce (Věra Schmiedtová) ... 9
Korpus a pravopis (Věra Schmiedtová) ... 11
Slavní předchůdci Českého národního korpusu I (Karel Kučera) ... 13
Trojúhelníky v korpusu (Marie Kopřivová) ... 14
Světelný rok (Jaroslava Hlaváčová) ... 16
Pokrývky hlavy (Věra Schmiedtová) ... 18
Země slovy korpusu (Marie Kopřivová) ... 19
Slavní předchůdci Českého národního korpusu II (Karel Kučera) ... 21
Kolokací, co hrdlo ráčí (Jana Klímová) ... 22
Členové rodiny v korpusu (Jana Klímová) ... 26

## POHLED NA JAZYK

Konference. Staré známé, nebo neznámé slovo? (František Čermák) ... 27
Jazyková poradna, korpus a internet (1): O Internetu a internetu (Josef Šimandl) ... 30
Kočkopes (Renata Blatná) ... 32
Jazyková poradna, korpus a internet (2): Rekupace (Josef Šimandl) ... 36
Slavní předchůdci Českého národního korpusu III (Karel Kučera) ... 37
Dialog mezi dvěma břehy (Renata Blatná a Jaroslava Hlaváčová) ... 38
Já vůl, ty vole/ty jsi vůl, to je vůl (František Čermák) ... 42
Jazyková poradna, korpus a internet (3): Acquis (Josef Šimandl) ... 45
Slavní předchůdci Českého národního korpusu IV (Karel Kučera) ... 47

## POHLED Z BOHEMISTICKÝCH ŠKAMEN

Kolísání rodu (Urszula Lisowska) ... 48
Slavní předchůdci Českého národního korpusu V (Karel Kučera) ... 50
Několik slov o slovech expresívních (Vendula Flaišmanová) ... 51
*Čert* a *ďábel* na materiálu ČNK (Hana Matysková) ... 53
Slovo *střecha* a jeho významy (Petra Smetáková) ... 54

Chodit do, chodit na (Miroslava Kazdová) ..... 56
Slavní předchůdci Českého národního korpusu VI (Karel Kučera) ..... 58

## KOLEGOVÉ A PŘÍZNIVCI

Český národní korpus a jeho jazykové značkování (Vladimír Petkevič) ..... 60
Slavní předchůdci Českého národního korpusu VII (Karel Kučera) ..... 74
Český národní korpus a lexikální archív Ústavu pro jazyk český v Praze (Věra Schmiedtová) ..... 75
Korpus a vývoj češtiny (Karel Kučera) ..... 78
Clara: Corpus Linguae Arabicae a problémy počítačového zpracování arabštiny (Petr Zemánek) ..... 82

## KORPUSOVÝ KVÍZ

Porovnejte své jazykové povědomí s korpusem (Josef Šimandl) ..... 90
Korpusový test (Marie Kopřivová a Michal Křen) ..... 91
Za jakých podmínek můžete pracovat s korpusem SYN2000? (Věra Schmiedtová) ..... 121